W | 200

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-122509
(43)Date of publication of application : 25.04.2003

(51)Int.Cl. G06F 3/06
G06F 12/00

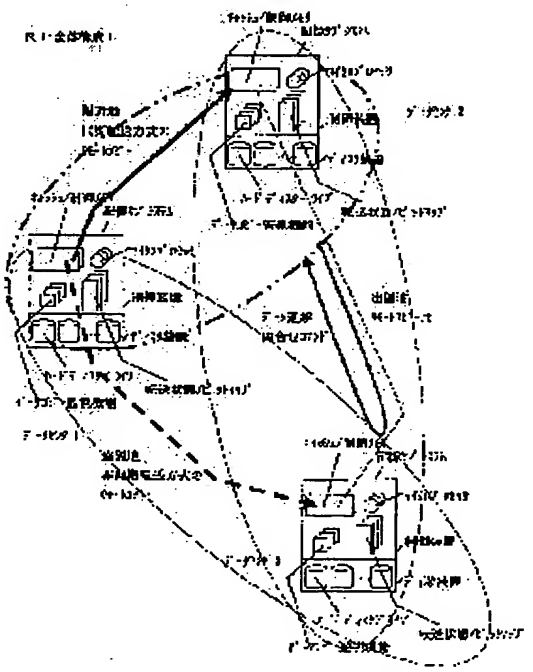
(21)Application number : 2002-019971
(22)Date of filing : 29.01.2002

(71)Applicant : HITACHI LTD
(72)Inventor : NAKANO TOSHIO
NAKAMURA KATSUNORI
OGATA MIKITO
OKAMI YOSHINORI
HIGAKI SEIICHI
ABEI MASARU
KIJIRO SHIGERU

(30)Priority
Priority number : 2001240072 Priority date : 08.08.2001 Priority country : JP

(54) REMOTE COPY CONTROL METHOD, STORAGE SUB-SYSTEM USING IT, AND WIDE AREA DATA STORAGE SYSTEM USING THEM

(57)Abstract:
PROBLEM TO BE SOLVED: To always hold the sequence of updating data between three or more data centers.
SOLUTION: Two data centers existing neighboring places are connected using a copy function by simultaneous transfer. One of the data centers and a third data center existing in a remote place are connected by an asynchronous remote copy function, whereby a storage sub-system existing in a neighboring place always ensures the sequence of the data received from a host and the third data center holds the data. Further, each storage sub-system is provided with a function of grasping the progress state of transferring, receiving and updating the data between the storage sub-systems installed in two data centers which do not directly transfer data in normal operation.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－122509
（P2003－122509A）
(43)公開日　平成15年４月25日(2003.4.25)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　3/06	３０４	Ｇ０６Ｆ　3/06	３０４Ｆ	５Ｂ０６５
12/00	５３１	12/00	５３１Ｄ	５Ｂ０８２
	５３３		５３３Ｊ	

審査請求　未請求　請求項の数34　ＯＬ　（全 34 頁）

(21)出願番号　特願2002－19971(P2002－19971)

(22)出願日　平成14年１月29日(2002.1.29)

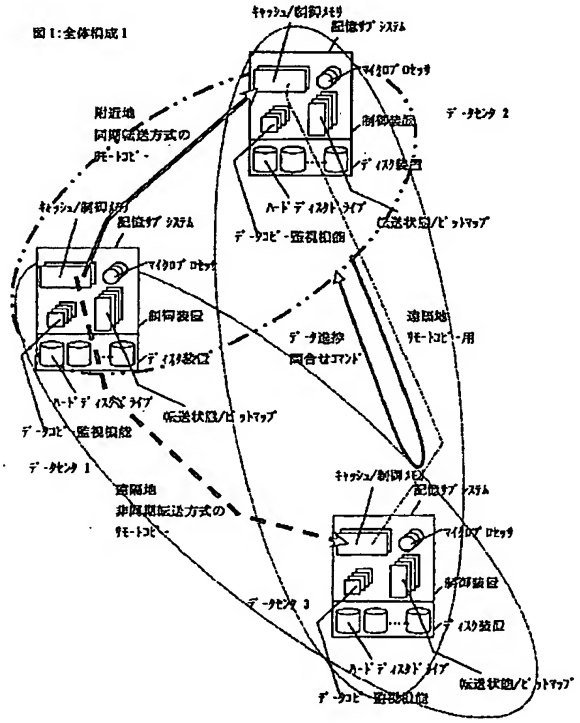
(31)優先権主張番号　特願2001－240072(P2001－240072)
(32)優先日　平成13年８月８日(2001.8.8)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　中野　俊夫
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　中村　勝寛
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(74)代理人　100071283
弁理士　一色　健輔

最終頁に続く

(54)【発明の名称】　リモートコピー制御方法、これを用いた記憶サブシステム、及び、これらを用いた広域データストレージシステム

(57)【要約】　（修正有）
【課題】３以上のデータセンタ間で、常時、データ更新の順序性を保持する。
【解決手段】　付近地に存在する２つのデータセンタ間は同期転送によるコピー機能を用いた接続構成とする。これらのうち１つのデータセンタと、遠隔地に存在する第３のデータセンタとの間は、非同期リモートコピー機能で連結し、付近地に存在する記憶サブシステムがホストから受領したデータの順序性を常時、保証しつつ第３のデータセンタが、そのデータを保持する。更に、正常運用の際には直接にデータ転送を行なわない２つのデータセンタに設置された記憶サブシステムの間で、データ転送・受領・更新の進捗状態を把握する機能を、各記憶サブシステムに設ける。



【特許請求の範囲】

【請求項１】　制御情報を格納する制御メモリと、データを一時的に格納するキャッシュメモリと、これらを制御するマイクロプロセッサにより、前記データを同期転送するＮ個の転送先と、前記データを非同期転送するＭ個の転送先とを有する記憶サブシステムのリモートコピー制御方法において、

前記制御メモリが、Ｎ＋Ｍ個のデータの転送先に対応する転送状態／ビットマップを格納する第１のステップと、

前記転送状態／ビットマップのうち１つに対応する転送状態／ビットマップを有する別の記憶サブシステムであって、直接にデータ転送を行っていないものに対して、データ更新の進捗状態を問合せるコマンドを発する第２のステップと、第２のステップの応答を受けて、転送状態／ビットマップを更新する第３のステップとを有することを特徴とするリモートコピー制御方法。

【請求項２】　請求項１記載のリモートコピー制御方法において、

第３のステップにおける転送状態／ビットマップの更新は、前記データ更新の進捗状態を問合せるコマンドが含む、データブロックの一部に対するデータの更新回数を計数するためのカウンタ値を更新することを含むリモートコピー制御方法。

【請求項３】　制御情報を格納する制御メモリと、データを一時的に格納するキャッシュメモリと、これらを制御するマイクロプロセッサにより、前記データを同期転送する転送先と、前記データを非同期転送する転送先とを有する記憶サブシステムにおいて、

直接データの転送を行わない別の記憶サブシステムに対し、データ更新の進捗状態を問合せる機能を有することを特徴とする記憶サブシステム。

【請求項４】　請求項３記載の記憶サブシステムは、更に、

自己と、直接データの転送を行わない別の記憶サブシステムにおいて保持する、１組の転送状態／ビットマップを有し、前記進捗状態を問合せる機能により、当該１組の転送状態／ビットマップを更新する記憶サブシステム。

【請求項５】　請求項４記載の記憶サブシステムにおいて、

前記１組の転送状態／ビットマップは、データブロックの一部に対するデータの更新回数を計数するためのカウンタ値を格納する領域を有する記憶サブシステム。

【請求項６】　第１のデータセンタに設置された第１の記憶サブシステム、第２のデータセンタに設置された第２の記憶サブシステム、及び、第３のデータセンタに設置された第３の記憶サブシステムを有する広域データストレージシステムにおいて、

第１の記憶サブシステムと第２の記憶サブシステムは、同期転送によりデータが転送され、

第１の記憶サブシステムと第３の記憶サブシステムは、非同期転送によりデータが転送され、第３の記憶サブシステムにおいて、当該転送されたデータの順序性が、常時、保たれることを特徴とする広域データストレージシステム。

【請求項７】　請求項６記載の広域データストレージシステムにおいて、

第２の記憶サブシステムから、同期転送によりデータの転送を受けた第１の記憶サブシステムが、第３の記憶サブシステムに対し、非同期転送により当該データを転送する広域データストレージシステム。

【請求項８】　請求項６記載の広域データストレージシステムにおいて、

第１の記憶サブシステムは、ホストから受領したデータを、同期転送により第２の記憶サブシステムへ転送すると共に、非同期転送により第３の記憶サブシステムにも転送する広域データストレージシステム。

【請求項９】　請求項６記載の広域データストレージシステムにおいて、

第１のデータセンタと第２のデータセンタは附近地に存在し、第１のデータセンタと第３のデータセンタは遠隔地に存在する広域データストレージシステム。

【請求項１０】　第１のデータセンタに設置された第１の記憶サブシステム、第２のデータセンタに設置された第２の記憶サブシステム、及び、第３のデータセンタに設置された第３の記憶サブシステムを有する広域データストレージシステムにおける、３つ以上のデータセンタの間のリモートコピー制御方法であって、

ホストからのデータを第１の記憶サブシステムが受領する第１のステップと、

第１の記憶サブシステムが、前記ホストからのデータを、第２の記憶サブシステムへ、同期転送する第２のステップと、

第１の記憶サブシステムが、前記ホストからのデータを、第３の記憶サブシステムへ、非同期転送する第３のステップと、

第２の記憶サブシステムから第３の記憶サブシステムへ、前記ホストからのデータが第３の記憶サブシステムへ到着したか否かを問合せる第４のステップとを有するリモートコピー制御方法。

【請求項１１】　請求項１０記載のリモートコピー制御方法において、更に、

第１のデータセンタが機能停止する第５のステップと、

第２の記憶サブシステムから第３の記憶サブシステムへ、第２の記憶サブシステムが保持するデータの一部を転送する第６のステップとを有するリモートコピー制御方法。

【請求項１２】　第１のデータセンタに設置された第１の記憶サブシステム、第２のデータセンタに設置された

第２の記憶サブシステム、及び、第３のデータセンタに設置された第３の記憶サブシステムを有する広域データストレージシステムにおける、３つ以上のデータセンタの間のリモートコピー制御方法であって、
ホストからのデータを第１の記憶サブシステムが受領する第１のステップと、
第１の記憶サブシステムが、前記ホストからのデータを、第２の記憶サブシステムへ、同期転送する第２のステップと、
第２の記憶サブシステムが、同期転送された前記ホストからのデータを、第３の記憶サブシステムへ、非同期転送する第３のステップと、
第１の記憶サブシステムから第３の記憶サブシステムへ、前記ホストからのデータが第３の記憶サブシステムへ到着したか否かを問合せる第４のステップとを有するリモートコピー制御方法。
【請求項１３】　請求項１２記載のリモートコピー制御方法において、更に、　第２のデータセンタが機能停止する第５のステップと、
第１の記憶サブシステムから第３の記憶サブシステムへ、第１の記憶サブシステムが保持するデータの一部を転送する第６のステップとを有するリモートコピー制御方法。
【請求項１４】　記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおけるリモートコピー制御方法において、
記憶サブシステムが、前記記憶資源にデータを書き込むステップと、
記憶サブシステムが、前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶するステップと、
記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先に送信するステップと、
記憶サブシステムが、前記転送先から送られてくる前記シーケンス番号を受信するステップと、
記憶サブシステムが、記憶している前記対応づけと前記転送先から受信した前記シーケンス番号とに基づいて、前記転送先において未反映となっている書き込みデータを把握するステップと、
を有することを特徴とするリモートコピー制御方法。
【請求項１５】　第１の記憶資源に対するデータ書き込み手段を備え第１のサイトに設置された第１の記憶サブシステムと、第１の記憶サブシステムに接続するホストと、第２の記憶資源に対するデータ書き込み手段を備え第２のサイトに設置された第２の記憶サブシステムと、第３の記憶資源に対するデータ書き込み手段を備え第３のサイトに設置された第３の記憶サブシステムと、を有する広域データストレージシステムにおけるリモートコピー制御方法において、
第１の記憶サブシステムが、前記ホストからの指示により第１の記憶資源に対してデータの書き込みを行うステップと、
第１の記憶サブシステムが、第１の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶するステップと、
第１の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを第２の記憶サブシステムに送信するステップと、
第２の記憶サブシステムが、前記データと前記対応づけとを受信して前記データを第２の記憶資源に記憶し、前記データと前記対応づけとを第３の記憶サブシステムに送信するステップと、
第３の記憶サブシステムが、前記データと前記対応づけとを受信して前記データを第３の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を第１の記憶サブシステムに送信するステップと、
第１の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、第３の記憶資源に未反映となっているデータを把握するステップと、
を有することを特徴とするリモートコピー制御方法。
【請求項１６】　前記第２の記憶サブシステムが障害等により使用できなくなった場合に、
前記第１の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第３の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第３の記憶資源に記憶して前記第１の記憶資源と前記第３の記憶資源の内容を同期させるステップと、
第１の記憶サブシステムが、前記ホストからの指示により前記第１の記憶資源に対してデータ書き込みが行われた場合に、前記データ書き込みにより書き込まれたデータと前記第１の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記データと前記書き込み位置情報とを受信して前記データを前記第３の記憶資源の前記書き込み位置情報で特定される位置に記憶するステップと、
を有することを特徴とする請求項１５に記載のリモートコピー制御方法。
【請求項１７】　前記第２の記憶サブシステムが回復し

再び使用できるようになった場合に、
前記第１の記憶サブシステムが、前記第１の記憶資源に格納されている全データを前記第２の記憶資源に転送するステップと、
第１の記憶サブシステムが、前記ホストからの指示により第１の記憶資源にデータ書き込みを行い、書き込んだデータとシーケンス番号とを第２の記憶サブシステムおよび第３の記憶サブシステムに送信するステップと、
第２の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを第２の記憶資源に記憶するステップと、
前記第２の記憶サブシステムが、前記第２の記憶資源に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶するステップと（位置情報管理テーブル作成）、
前記第３の記憶サブシステムが、記憶サブシステム１から送られてくる前記データと前記シーケンス番号とを受信して前記データを前記第３の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を前記第２の記憶サブシステムに送信するステップと、
前記第１の記憶サブシステムから前記第３の記憶サブシステムへのデータの転送を停止するステップと、
前記第２の記憶サブシステムが、記憶サブシステム３から送られてくる前記シーケンス番号を受信して、前記シーケンス番号と自身が記憶している前記シーケンス番号とこれに対応する書き込み位置情報とに基づいて、前記第３の記憶資源に未反映となっているデータを把握するステップと、
前記第２の記憶サブシステムが、前記把握した前記第３の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第３の記憶資源に記憶して前記第１の記憶資源と前記第３の記憶資源の内容を同期させるステップと、
を有することを特徴とする請求項１６に記載のリモートコピー制御方法。
【請求項１８】　記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおけるリモートコピー制御方法において、
記憶サブシステムが、前記記憶資源にデータを書き込むステップと、
記憶サブシステムが、前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを生成するステップと、
記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先の一つである転送先Ａに送信するステップと、
転送先Ａが記憶サブシステムから送信されてくる前記対応づけを受信して記憶するステップと、
転送先Ａが別の前記転送先である転送先Ｂから送られてくる前記シーケンス番号を受信するステップと、
転送先Ａが、記憶している前記対応づけと転送先Ｂから受信した前記シーケンス番号とに基づいて、転送先Ｂにおいて未反映となっている書き込みデータを把握するステップと、
を有することを特徴とするリモートコピー制御方法。
【請求項１９】　第１の記憶資源に対するデータ書き込み手段を備え第１のサイトに設置された第１の記憶サブシステムと、第２の記憶資源に対するデータ書き込み手段を備え第２のサイトに設置された第２の記憶サブシステムと、第２の記憶サブシステムに接続するホストと、第３の記憶資源に対するデータ書き込み手段を備え第３のサイトに設置された第３の記憶サブシステムと、を有する広域データストレージシステムにおけるリモートコピー制御方法であって、
第２の記憶サブシステムが、前記ホストからの指示により第２の記憶資源に対してデータの書き込みを行うステップと、
第２の記憶サブシステムが、前記書き込んだデータと、前記書き込みが行われた位置とを特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号とを第１の記憶サブシステムに送信するステップと、
第１の記憶サブシステムが前記データと前記シーケンス番号とを受信してこれらを第１の記憶資源に記憶し、前記シーケンス番号と、前記データが書き込まれた第１の記憶資源上の格納位置を特定する書き込み位置情報とを対応づけて記憶するステップと、
第２の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記シーケンス番号とを第３の記憶サブシステムに送信するステップと、
第３の記憶サブシステムが前記データと前記シーケンス番号とを受信して前記データを第３の記憶資源に記憶するとともに前記データと対になっていた前記シーケンス番号を第１の記憶サブシステムに送信するステップと、
第１の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、第３の記憶資源に未反映となっているデータを把握するステップと、
を有することを特徴とするリモートコピー制御方法。
【請求項２０】　前記第２の記憶サブシステムが障害等により使用できなくなった場合に、
前記第１の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第３の記憶資源において未反映となっている差分のデータ

とその書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第３の記憶資源に記憶して前記第１の記憶資源と前記第３の記憶資源の内容を同期させるステップと、
第１の記憶サブシステムが、前記ホストからの指示により前記第１の記憶資源に対してデータ書き込みが行われた場合に、前記データ書き込みにより書き込まれたデータと前記第１の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記データと前記書き込み位置情報とを受信して前記データを前記第３の記憶資源の前記書き込み位置情報で特定される位置に記憶するステップと、を有することを特徴とする請求項１９に記載のリモートコピー制御方法。
【請求項２１】　前記第２の記憶サブシステムが回復し再び使用できるようになった場合に、
前記第１の記憶サブシステムが、前記ホストにより書き込まれた前記第１の記憶資源に格納されている全てのデータを前記第２の記憶資源に転送するステップと、
第１の記憶サブシステムが、前記ホストからの指示により第１の記憶資源に対してデータ書き込みを行うステップと、
第１の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記シーケンス番号とを第２の記憶サブシステムおよび第３の記憶サブシステムに送信するステップと、
第２の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを第２の記憶資源に適時に記憶するステップと、
前記第２の記憶サブシステムが、前記第２の記憶資源に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶するステップと、
前記第３の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを前記第３の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を前記第２の記憶サブシステムに送信するステップと、
前記第１の記憶サブシステムから前記第２の記憶サブシステムへのデータの転送を停止するステップと、
前記第２の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、前記第３の記憶資源に未反映となっているデータを把握するステップと、
前記第２の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第３の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第３の記憶サブシステムに送信するステップと、
前記第３の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第３の記憶資源に記憶して前記第１の記憶資源と前記第３の記憶資源の内容を同期させるステップと、
を有することを特徴とする請求項２０に記載のリモートコピー制御方法。
【請求項２２】　前記第１の記憶サブシステムと前記第２の記憶サブシステムとの間は同期転送により運用され、前記第１の記憶サブシステムと前記第３の記憶サブシステムとの間は非同期転送により運用されることを特徴とする請求項１５～１７および１９～２１のいずれかに記載のリモートコピー制御方法。
【請求項２３】　前記第２の記憶サブシステムもしくは前記第３の記憶サブシステムは、前記第１の記憶サブシステムにおける書き込み位置情報に対応づけた前記第２の記憶資源もしくは前記第３の記憶資源における格納位置にデータを記憶することを特徴とする、請求項１５～１７および１９～２１のいずれかに記載のリモートコピー制御方法。
【請求項２４】　ホストが接続された記憶サブシステムＡとこれに通信手段を介して接続する複数の他の記憶サブシステムＢと、を有する広域データストレージシステムにおけるリモートコピー制御方法であって、
記憶サブシステムＢが、前記ホスト、記憶サブシステムＡ、前記通信手段のうち少なくともいずれか一つに障害が発生したことを検知するステップと、
記憶サブシステムＢが、障害を検知した場合に記憶サブシステムＢの中から記憶サブシステムＡの処理を代行させる記憶サブシステムｂを選出するステップと、
記憶サブシステムＢが、選出した記憶サブシステムｂとこれ以外の記憶サブシステムＢが記憶しているデータの内容を一致させるステップと、記憶サブシステムＢが、記憶サブシステムｂに接続する副ホストにより前記ホストの運用を引き継ぐステップと、
を備えることを特徴とするリモートコピー制御方法。
【請求項２５】　障害が発生したことを検知する前記ステップ、および記憶サブシステムＡの処理を代行させる記憶サブシステムｂを選出する前記ステップを、一台以上の前記記憶サブシステムＢが行うことを特徴とする請求項２４に記載のリモートコピー制御方法。
【請求項２６】　前記障害が発生したことを検知するステップが、
前記記憶サブシステムＡから前記記憶サブシステムＢに対して送信されるハートビートメッセージがあらかじめ設定された時刻に受信できない場合に障害が発生したと認識するステップであることを特徴とする請求項２４に

記載のリモートコピー制御方法。

【請求項２７】　障害を検知した場合に記憶サブシステムＢの中から記憶サブシステムＡの処理を代行させる記憶サブシステムｂを選出する前記ステップが、

前記各記憶サブシステムＢに記憶しているデータの更新順序を示すシーケンス番号のうち最新のシーケンス番号同士を比較して、これらのうちで最新の前記シーケンス番号を記憶している記憶サブシステムＢを、前記記憶サブシステムＡの処理を代行させる前記記憶サブシステムｂとして選出するステップであることを特徴とする請求項２４に記載のリモートコピー制御方法。

【請求項２８】　前記各記憶サブシステムＢに記憶しているデータの更新順序を示す前記シーケンス番号に抜けがある場合に、連続しているシーケンス番号のうち最新のシーケンス番号を用いて前記比較を行うことを特徴とする請求項２７に記載のリモートコピー制御方法。

【請求項２９】　選出した記憶サブシステムｂとこれ以外の記憶サブシステムＢが記憶しているデータの内容を一致させる前記ステップが、

前記記憶サブシステムｂとこれ以外の記憶サブシステムとの間で、全データの複写、もしくは、各記憶サブシステムＢに記憶しているデータの差分データの複写により、前記選出した記憶サブシステムｂとこれ以外の記憶サブシステムＢが記憶しているデータの内容を一致させるステップであることを特徴とする請求項２４に記載のリモートコピー制御方法。

【請求項３０】　記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおける請求項１４に記載のリモートコピー制御方法に使用する前記記憶サブシステムであって、

前記記憶資源に対してデータ書き込みを行う手段と、

前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶する手段と、

前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先に送信する手段と、前記転送先から送られてくる前記シーケンス番号を受信する手段と、

記憶している前記対応づけと前記転送先から受信した前記シーケンス番号とに基づいて、前記転送先において未反映となっている書き込みデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項３１】　第１の記憶資源に対するデータ書き込み手段を備え第１のサイトに設置された第１の記憶サブシステムと、第１の記憶サブシステムに接続するホストと、第２の記憶資源に対するデータ書き込み手段を備え第２のサイトに設置された第２の記憶サブシステムと、第３の記憶資源に対するデータ書き込み手段を備え第３のサイトに設置された第３の記憶サブシステムと、を有する広域データストレージシステムにおける請求項１５に記載のリモートコピー制御方法に使用する前記第１の記憶サブシステムとして機能する記憶サブシステムであって、

前記ホストからの指示により第１の記憶資源に対してデータの書き込みを行う手段と、

第１の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶する手段と、

前記データ書き込みにより書き込んだデータと前記対応づけとを第２の記憶サブシステムに送信する手段と、第３の記憶サブシステムから送信されてくるシーケンス番号を受信して、このシーケンス番号と記憶している前記対応づけとに基づいて、第３の記憶資源に未反映となっているデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項３２】　記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムＡにおける請求項１８に記載のリモートコピー制御方法に使用する前記転送先として機能する記憶サブシステムであって、

記憶サブシステムＡから送られてくる、前記記憶資源に対して書き込まれたデータと、そのデータの書き込みが行われた位置を特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号との対応づけとを受信して記憶する手段と、

別の前記転送先である転送先Ｂから送られてくる前記シーケンス番号を受信する手段と、

記憶している前記対応づけと転送先Ｂから受信した前記シーケンス番号とに基づいて、転送先Ｂにおいて未反映となっている書き込みデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項３３】　第１の記憶資源に対するデータ書き込み手段を備え第１のサイトに設置された第１の記憶サブシステムと、第２の記憶資源に対するデータ書き込み手段を備え第２のサイトに設置された第２の記憶サブシステムと、第２の記憶サブシステムに接続するホストと、第３の記憶資源に対するデータ書き込み手段を備え第３のサイトに設置された第３の記憶サブシステムと、を有する広域データストレージシステムにおける請求項１９に記載のリモートコピー制御方法に用いられる前記第１の記憶サブシステムとして機能する記憶サブシステムであって、

第２の記憶サブシステムから送信されてくる、前記ホストからの指示により書き込んだデータと、前記書き込みが行われた位置とを特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号との対応づけとを受信して、これを第１の記憶資源に記憶する手段と、

第３の記憶サブシステムから送られてくるシーケンス番号を受信して、このシーケンス番号と、自身が記憶している前記対応づけとに基づいて、第３の記憶資源に未反映となっているデータを把握する手段と、
を備えることを特徴とする記憶サブシステム。
【請求項３４】　ホストが接続された記憶サブシステムＡとこれに通信手段を介して接続する複数の他の記憶サブシステムＢと、を有する広域データストレージシステムにおける請求項２４に記載のリモートコピー制御方法に使用する前記記憶サブシステムＢとして機能する記憶サブシステムであって、
前記ホスト、記憶サブシステムＡ、前記通信手段のうち少なくともいずれか一
つに障害が発生したことを検知する手段と、
障害を検知した場合に記憶サブシステムＢの中から記憶サブシステムＡの処理を代行させる記憶サブシステムｂを選出する手段と、
選出した記憶サブシステムｂとこれ以外の記憶サブシステムＢが記憶しているデータの内容を一致させる手段と、
記憶サブシステムｂに接続する副ホストにより前記ホストの運用を引き継ぐ手段と、
を備えることを特徴とする記憶サブシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、災害による外部記憶装置の障害が生じた後に、速やかに、その障害から復旧可能な広域データストレージシステムに係り、特に、外部記憶装置が相互に１００ｋｍから数百ｋｍ隔てて設置され、相補的な動作を行う３つ以上の外部記憶装置からなる広域データストレージシステムに関する。
【０００２】
【従来の技術】本件の出願人による特開平１１－３３８６４７号公報には、システムとデータの２重化を同期又は非同期にて行うことが開示されている。また、本件の出願人による特開２０００－３０５８５６号公報には、非同期で遠隔地にデータのコピーを行う技術が開示されている。
【０００３】このように、本件の出願人は、大型計算機システム、サーバー、ネットワーク上のパーソナルコンピュータ、その他の上位計算機システム（以下、ホストという。）から、データの順序を特定する特別な制御情報を受領することなく、そのデータを受け取った外部記憶装置（以下、記憶サブシステムという。）が、そのデータを、遠隔地に設置された第２の記憶サブシステムに対し、そのデータの順序性を常時保証しながら、非同期転送により第２の記憶サブシステムへ連続して間断なく書き込むという、非同期リモートコピーの技術を所有している。
【０００４】また、同期転送の技術を用いてコピーを行うときは、ホストとこれに接続された記憶サブシステムとの間のデータ更新処理と、この記憶サブシステムと付近地又は遠隔地に設置された記憶サブシステムとの間のコピー制御が連動するため、巨視的にみて常に２つの記憶サブシステム間でデータが一致しており、その書込み順序性も同時に保証されている。尚、適当なデータ転送経路を選択すれば、２つの記憶サブシステムの距離が１００ｋｍを超える場合であっても、同期転送によるコピーが可能である。
【０００５】昨今、データを安全に格納し保持することが重要であるという認識が高まっており、データストレージの市場では、ディザスタリカバリシステムを要請する声が多く聞かれる。従来のように、データ格納拠点を２つ設け、かかる２地点間を同期転送又は非同期転送で結ぶことは実現されている。しかし市場は、第３、第４のデータ格納拠点（以下、データセンタという。）を要求し、これらの間での完全な又は完全に近いディザスタリカバリシステムの構築を望んでいる。
【０００６】その理由は、３拠点以上のデータセンタを設置しておけば、これらのうち１個所が災害に見舞われても、引き続き発生する災害のリスクを軽減するために、残る複数のセンタ間でデータ冗長化の回復・維持が図れるであろうという期待にある。
【０００７】従来の技術では、３以上のデータセンタを構築した場合に、ホストから受領するＩ／Ｏを唯一の記憶サブシステムの論理ボリュームで受領し、これを複数のデータセンタへリモートコピー技術を用いて転送する際の配慮が十分で無かった。例えば、一つのデータセンタが災害によりダウンした場合に、残る２以上のデータセンタ間で、データの順序性を保証した論理ボリュームを構築できるか、更新状態を引き継ぎデータの不整合を無くすことができるか、附近地又は遠隔地に対するコピーを可能とするシステムの再構築ができるかといった問題に関し、配慮が足りなかった。
【０００８】
【発明が解決しようとする課題】災害はいつ発生するか不明なため、３以上のデータセンタ間で、常時、データ更新の順序性を保持しなければならない。
【０００９】このため、ホストに特殊な機能を具備せず、複数のリモートコピー構成を連結し、同一論理ボリュームが受領したデータを、遠隔地又は附近地の別の記憶サブシステムへ配信し、かつ、如何なる時点で災害が発生しても、ホストからのデータの更新順序を、各データセンタの記憶サブシステムで、常時、保証する広域データストレージシステムを構成しなければならない。
【００１０】本発明に係る広域データストレージシステムでは、記憶サブシステムの内部に、冗長化した論理ボリュームを設けることなく、別の記憶サブシステムに対し、データをコピーすることにより上記の課題を解決している。

【００１１】また、本発明に係る広域データストレージシステムでは、災害後の復旧作業として、広域ストレージシステムの再構成を想定しており、正常な運用時に、直接、データ転送を行っていない記憶サブシステム間で、管理情報を遣り取りし、データの更新状態を各記憶サブシステムで監視し管理する。そして、災害後の復旧作業（再同期、リシンク）において、災害発生直前に各記憶サブシステムが保持しているデータの差分のみを転送することで、即時に、ホストの交代（ｆａｉｌｏｖｅｒ）と、アプリケーション実行の継続を行う。

【００１２】＜データ更新の順序性を常時保証することについて＞ここで、順序性の保持の時間的範囲について補足説明する。

【００１３】ホストから発行されたＩ／Ｏは記憶サブシステムに書き込まれ、記憶サブシステムが報告するデータの書き込み完了報告を認識し、ホストは次のステップを実行する。ホストは記憶サブシステムのデータ書き込み完了を受領しない場合又は障害報告があった場合は、次のＩ／Ｏを正常には発行しない。従ってデータの書きこみの順序性は、ホストが記憶サブシステムから書き込み完了報告を受領する前後で、記憶サブシステムが順序性保存の何らかの処理をすることで維持されるべきものである。

【００１４】同期転送のリモートコピーでは転送されコピーされるデータが付近地又は遠隔地（以下、単に「別地」と略記する。）の記憶サブシステムに書き込まれ、別地の記憶サブシステムからの書き込み完了を受領した後、ホストに対し書き込み完了報告を行う。リモートコピーを行わない場合と比較し、リモートコピーに係る処理、及びデータ転送処理時間が長くかかり、性能が遅延する。リモートコピーにおける接続距離を延長すると、データ転送に伴う処理時間が増大し、リモートコピーを行うことによりホストのＩ／Ｏ処理の性能をさらに低下させる。これを打破する一つの方法が非同期転送である。

【００１５】非同期転送は、ホストからＩ／Ｏを受領した記憶サブシステムが、別地の記憶サブシステムへのデータ転送を行ない別地の記憶サブシステムの書き込み完了を待たずに、ホストからＩ／Ｏを受領した記憶サブシステムが書き込み完了報告をホストへ返す。これにより、記憶装置サブシステム間のデータ転送は、ホストのＩ／Ｏ処理と関係が無くなり、ホストのＩ／Ｏ処理と非同期に実行できる。しかし、ホストからのデータの到着順序を守って、別地の記憶サブシステムへデータを書き込まなければ、別地の記憶サブシステムのデータ順序性は維持されず、両記憶サブシステム間でデータの不整合を来す可能性がある。データの順序性を常時保証する機能を追加すれば、このような可能性を極小化できる。

【００１６】別地の記憶サブシステムは、ホストＩ／Ｏを受領した記憶サブシステムと比較し、通常はデータの更新は遅れているが、ホストからのデータ到着順序を守って記憶サブシステムへ書き込む限り、データの順序性に矛盾は無く、ジャーナルファイルシステムやデータベースリカバリ処理により、障害時の回復が可能である。

【００１７】一方、データの順序性を維持せず、別地の記憶サブシステムへリモートコピーしてデータを反映させる方法もある。この方法は、ある時点までのホストから受領したデータを別地へ送り、それらを記憶サブシステムへ纏め書きする。ある時点までのデータ書き込みが終わった段階で、データ転送を終了し、以降、次回の纏め書きまで、リモートコピーのデータ転送を抑止し、抑止している間のデータ順序性、ホストから受領したＩ／Ｏの一貫性を保証する。

【００１８】この方法では、データの順序情報を付与する機能が不要であるが、ある程度の更新分のデータを蓄えておいて、その更新分を一括転送し、リモートへ書き込みが全て完了した段階で、データ整合性を保証している。この方法ではリモートコピーを行っている間に障害が発生すると、リモート側のデータ更新は順序性を維持して更新されていないため全滅となる恐れがある。リモートコピーのデータ転送を止めている間のみ、データ整合性を保証でき、adaptiveと呼ばれている。

【００１９】出願人の所有する"データの順序性を常時保証する非同期転送によるリモートコピー"の技術によれば、ホストに完了報告を返す際に、記憶サブシステムがデータの順序性を保証する処理をしていることが特徴である。記憶サブシステムの制御装置におけるオーバヘッドや内部処理の遅延時間に拘らず、ホストに返す際にデータ順序情報をブロック毎に管理する措置を施しているため、常時順序性を保証できる。

【００２０】実際には、ホストから受領するＩ／Ｏ発行間隔よりかなり短い時間で、ブロック毎の制御・管理をしている。この一方で、リモート側の記憶サブシステムでデータ配信を待ちきれずタイムアウト（Timeout）とする値は、１時間以上に設定可能でもある。大切なのは、出願人のリモートコピーの技術が、データに順序情報を付与してデータブロックを転送し、これに基づきデータの順序を守って書き込みを行なっている点である。ローカル／リモートのデータ更新の時間差が、例え半日あっても、順序性さえ正しければ、更新データ全てを喪失してしまう不整合より良いからである。

【００２１】

【課題を解決するための手段】データを同期及び非同期に転送可能な転送経路、所定の管理情報の遣り取りが可能な通信線路、及び、データ更新進捗管理手段により、３以上のデータセンタを相互に連結する構成とする。

【００２２】ここで、更新進捗管理手段は、各記憶サブシステムに設けられ、いつ発生するか分からない災害に対応するため、他のデータセンタに設置された記憶サブシステムにおけるデータ更新の進捗状態を、適宜、監視

し、相互にその記憶サブシステムのデータ更新状態を把握させる手段である。

【００２３】具体的には、直接データ転送を行なっていない記憶サブシステムの各々が、転送状態／ビットマップを持ち、転送ブロックのどの位置が何回更新されたか、一方が問合せ、他方がこれに応答することで、データ更新（リモートコピー）の進捗を監視し管理する機能を有する。

【００２４】

【発明の実施の形態】３以上のデータセンタに、それぞれ設置された記憶サブシステムの間を、同期転送により連結する。

【００２５】かつ、データの順序性を常時、連続的に保証する非同期リモートコピーの技術で連結する。そして、１箇所のプライマリデータセンタの記憶サブシステムから、これを除いた残りの別拠点の２以上のデータセンタの各記憶サブシステムへ、ホストからプライマリの記憶サブシステムが受領したデータを、ホストが更新した順序を保持しつつ、連続的に転送し格納する。

【００２６】データが、ホストからの更新順を保証して冗長構成化されるため、万一、データセンタに災害・障害が発生しても、残ったデータセンタの記憶サブシステムの間で、各記憶サブシステム間の差分データのみを転送することで、即時に、リモートコピーの運用構成を回復でき、又は、データ喪失を最小限度とすることができる。

【００２７】＜同期・非同期について＞まず始めに、図５、図６を用いて同期転送によるコピー又は非同期リモートコピーを定義する。

【００２８】同期転送によるコピーとは、ホスト１から記憶サブシステム１に、データの更新(書込み)指示が有った場合に、その指示対象が附近地に設置された記憶サブシステム２にも書込むデータであるときは、記憶サブシステム２に対して、指示された更新（書込み）が終了してから、ホストに更新処理の完了を報告する処理手順をいう。ここで、附近地とは、いわゆるメトロポリタンネットワークと称される１００ｋｍ程度までの範囲を言うものとする。

【００２９】つまり、同期転送のリモートコピー（図５）では、ホスト１から受領した更新データブロックを記憶サブシステム１で受領し（①）、そのデータブロックを記憶サブシステム２に転送し（②）、書き込み完了後、これを記憶サブシステム１で受領し（③）、最後にホスト１に対し更新データブロックの書き込み完了を行う（④）。途中の処理に失敗した場合には、ホスト１に書き込み障害を報告する。

【００３０】同期転送によるコピーを実施すると、ホスト１に接続された近い方の記憶サブシステム１と、附近地に設置された遠い方の記憶サブシステム２のデータの内容が、巨視的にみて常に一致している。このため、災害により一方がその機能を失った場合であっても、他方の記憶サブシステムに災害直前までの状態が完全に保存されているので、残るシステムで迅速に処理を再開できる効果がある。尚、巨視的にみて常に一致とは、同期転送の機能を実施中には、制御装置や電子回路の処理時間の単位（μsec、msec）で、一致していない状態が有り得るが、データ更新処理完了の時点ではデータは必ず同一の状態になっていることを意味している。これは、附近地の記憶サブシステムへの更新データの反映が終了しない限り、ホストに近い側の記憶サブシステム１の更新処理を完了できないためである。

【００３１】一方、非同期リモートコピー（図６）とは、ホスト１からこれに接続された近い方の記憶サブシステム１に、データの更新(書込み)指示が有った場合、その指示対象が遠隔地に設置された記憶サブシステム２にも書込むデータであっても、記憶サブシステム１の更新処理が終わり次第、ホスト１に対し更新処理の完了を報告し、遠隔地の記憶サブシステム２におけるデータの更新（反映）が、近い方の記憶サブシステム１における処理とは非同期に実行される処理手順をいう。

【００３２】このため、近い方の記憶サブシステム１で必要とされる処理時間でデータ更新が終了するので、遠隔地の記憶サブシステム２へのデータの格納に起因する伝送時間、格納処理時間等により、ホスト１の処理が待たされることがない。ここで、遠隔地とは、いわゆるトランスコンチネンタルネットワークと称される、附近地より遠いが、距離の制約なく通信又はデータ転送可能な地点を言うものとする。

【００３３】より具体的には、非同期リモートコピーでは、ホスト１から受領した更新データブロックを記憶サブシステム１で受領し（①）、ホスト１に対し更新データブロックの書き込み完了を行う（②）。記憶サブシステム１は、自己のスケジュールで、ホスト１の処理とは非同期に、記憶サブシステム２へデータを転送する。

【００３４】遠隔地又は附近地へのデータ転送経路の複雑化、途中のデータ転送経路のボトルネックにより、データ転送中の当該データの順序性は保証されない（図６、点線の楕円内参照）。

【００３５】一般に、データ転送の性能を上げるため、多くは高速転送のため、転送元から複数の転送経路を用いてデータを転送する場合がある。また、転送先まで遠距離となると、転送元は１つの転送経路であっても介在する交換機、ルーターその他の通信中継機器により、転送先まで転送経路が１本であることは保証されない。このように複数の転送経路を用いてデータを転送する場合には、経路によっては時間的な差異が生じ、遅い経路と早い経路とを介してデータが送られるため、転送先においてデータの順序が保存されないのである。

【００３６】図６の楕円内に一例を示すが、データ転送経路上の順序を、Ｄａｔａ＃１、Ｄａｔａ＃２、Ｄａｔ

ａ＃４、Ｄａｔａ＃３としている。記憶サブシステム２における更新順序はＤａｔａ＃１、Ｄａｔａ＃２、Ｄａｔａ＃３、Ｄａｔａ＃４の順序である。記憶サブシステム２において、転送されてきたデータの順序をソートして正規の順序に並べ直しているからである。この更新処理の直後に不慮の災害が発生しても、データ更新の順序が守られているため、記憶サブシステム２のデータベースやジャーナルファイルシステムは回復処理を行うことができる。逆に、更新処理の直前に災害が発生したときは回復処理は不可能であるが、ホストへの応答とは非同期に、記憶サブシステム間で連続的に間断無くデータ転送処理を行うことで、データ不整合を極小化でき、巨視的に見て、常時、更新データの順序性を確保できる。

【００３７】本実施の態様では、ホスト１からデータブロックを受領し、記憶サブシステム２へ転送する際に、ホストからのデータ更新順序を示すシーケンス番号情報をデータに付して転送している。このため、記憶サブシステム２で、シーケンス番号情報に基づくソート制御を行い、順序性を保証して、データの格納を完了できる。このような一連のデータ転送・処理に必要な処理時間の後は、データの順序性が遠隔地の記憶サブシステム２において保持されている。このように非同期のコピーを、これに固有なデータ処理を連続して行うこと（非同期リモートコピー）で、常時、データ更新の順序性を保証することができる。

【００３８】非同期リモートコピーは、ホスト１の処理性能を落とさず、記憶サブシステム１及び２の間の距離を拡大できる特長があり、かつ、常時、データの順序性が保証されるため、広域データストレージシステムの利用者が自己の業務を遂行する上で、ほぼ任意の時点のデータベースやジャーナルファイルシステムの整合性を、遠隔地に設置された記憶サブシステムにおいて確保できる特長を有している。

【００３９】＜広域データストレージシステム、その１＞図１に本発明の広域データストレージシステムの全体構成を示す。図９は本発明の別の広域データストレージシステムの全体構成を示す図である。図１０は図１と図９の構成の組み合わせによる応用例を示す図である。

【００４０】図１において３ヵ所のデータセンタに記憶サブシステムを設置する。各データセンタには複数の記憶サブシステムが設置されても良いし、それらがリモートコピー機能を伴った接続形態となっていても良い。アプリケーションはデータセンタ１に接続されたホストで実行される。尚、ホストとデータセンタ１とのデータ転送経路は、ファイバーチャネル、メインフレームインタフェース、イーサネット（登録商標）ＬＡＮ、公衆回線、インターネットその他専用回線である。

【００４１】データセンタ１とデータセンタ２は附近地に存在し、同期転送によりデータ転送し得る構成である。データセンタ１とデータセンタ３は遠隔地に存在し、これらの間は非同期リモートコピーの技術によりデータ転送し得る構成である。

【００４２】正常な運用形態では、ホストからデータセンタ１が受領した更新データは、データセンタ１に設置された記憶サブシステムに格納され運用される。この更新データは、附近地に設置されたデータセンタ２の記憶サブシステムへ、ファイバーチャネル、メインフレームインタフェース、イーサネットＬＡＮ、公衆回線、インターネットその他専用回線を介して、同期転送される。つまり、データセンタ１とデータセンタ２では、記憶サブシステム間のデータ整合性は巨視的には絶えず保たれている。

【００４３】正常な運用形態では、また、ホストからデータセンタ１が受領した上記の更新データを、遠隔地に設置されたデータセンタ３の記憶サブシステムへ、上記と同様な専用回線を介して、上記の同期転送処理と同時に、非同期リモートコピーの技術で転送される。尚、データセンタ１とデータセンタ２、データセンタ１とデータセンタ３、それぞれの間のデータ転送経路は同一種類の回線にする必要はない。

【００４４】データセンタ１とデータセンタ３との間は遠距離であり、この間の転送に起因する更新データの到着順序の不整合が生じる。また、転送元となるデータセンタ１の記憶サブシステムには、転送先で未反映のデータとなる差分データが、存在することとなる。しかし、本発明の非同期リモートコピーでは、所定の非同期転送に固有のデータ処理後は、データベースやファイルシステムの回復処理に必要な、ホストからのデータ順序性を保証しているために、不整合を生じていたデータの順序を回復させることが可能である。この結果、データセンタ１とデータセンタ３の記憶サブシステム間では、上記ホストから受領した更新データの順序性は保たれる。

【００４５】データセンタ２とデータセンタ３の間は、万一のリカバリ処理に備え、データを転送する通信線路は敷設・準備されているが、このストレージシステムの正常な運用時にはホストからの更新データは転送されない。データセンタ１での災害・障害発生の際に備え、正常な運用形態で、データ転送の進捗状態を問合せるコマンドが、この通信線路を介して、データセンタ２からデータセンタ３へ、又は逆にデータセンタ３からデータセンタ２へ、送受信されることとなる。尚、敷設・準備された通信線路は、ファイバーチャネル、メインフレームインタフェース、イーサネットＬＡＮ、公衆回線、インターネットその他専用回線である。

【００４６】正常時には、記憶サブシステム１と記憶サブシステム３との間の非同期リモートコピーにより為されたホストからの更新データの到着を、記憶サブシステム２から発せられる"データ進捗問合せコマンド"により、データセンタ２とデータセンタ３の間の通信線路を介して、問合せる。

【００４７】"データ進捗問合せコマンド"の起動は、記憶サブシステム２のスケジュールに従って為される。記憶サブシステム１からの同期転送によるデータの受領のタイミングで、当該コマンドを発行しても良いし、所定の時間間隔で纏めて問合せても良い。所定の時間間隔としては、例えば１００ｍｓｅｃから５００ｍｓｅｃ毎に問合せても良いが、後述する転送状態／ビットマップの管理、これに基づく差分データの管理に時間が費やされ過ぎない程度となる。尚、１回の問合せで複数のビットマップを検査するようにしても良い。

【００４８】正常な運用時には、記憶サブシステム２と記憶サブシステム３との間で、直接、データの転送は行なわれない。このため、記憶サブシステム２が"データ進捗問合せコマンド"を発行して、記憶サブシステム１と記憶サブシステム３のデータ更新状況を把握する。

【００４９】万一、データセンタ１で障害が発生したときには、データセンタ２のホストを用いて、これまでのシステム運用を続行し（ホストのフェールオーバー）、記憶サブシステム２と記憶サブシステム３との間の差分データを、リカバリ処理に備えて敷設されたデータ転送の通信線路を用いてデータセンタ２からデータセンタ３へ転送する。差分データのみの転送で即時に広域データストレージシステムを回復させることが可能である。
尚、フェールオーバーとは、プライマリーシステムからサブシステムへ切り替えることをいい、古くは、ホットスタンバイとも呼ばれていた。

【００５０】この後、データセンタ２からデータセンタ３へ、上記の通信線路を用いて、上述のような非同期リモートコピーを行うこととすれば、データセンタ１の復旧に伴う、データセンタ２とデータセンタ１との間の同期転送の復旧により、障害発生前の広域データストレージシステムを復旧させることがでる。但し、障害発生前後で、データセンタ１とデータセンタ２の役割が入れ替わっている。

【００５１】このように、附近地に存在する２つのデータセンタと、遠隔地に存在する２つのデータセンタとを統合し合計３つのデータセンタとすることで、リモートコピーの技術で連結する広域データストレージシステムとする。こうしておけば、中小規模の災害・障害のときは、附近地に存在する、相互に同期転送により連結されたデータセンタの一方で他方の代替を行うことができる。２つのデータセンタの記憶サブシステムのデータは同期転送により巨視的にみて一致しており、フェールオーバーが即時に行なえるからである。

【００５２】＜広域データストレージシステム、その２＞図１のデータセンタ２とデータセンタ３との間の通信線路が非常用であるため、この通信線路を選択せず、障害・災害復旧後のデータセンタ１とデータセンタ３との間のデータ転送経路を選択する場合には、復旧後は、広域データストレージシステムは、図９の構成となる。

【００５３】図９は、記憶サブシステム１と記憶サブシステム２が同期転送で、記憶サブシステム２と記憶サブシステム３が非同期リモートコピーで、それぞれ、接続された例である。図１の広域データストレージシステムにおいて、データセンタ１からデータセンタ２へ運用を切り替え、データセンタ２を主たる運用サイトとし、災害・障害復旧後は、データセンタ２からデータセンタ１へデータを同期転送させる一方で、データセンタ１からデータセンタ３へデータを非同期転送させる構成となるからである。

【００５４】図９の場合において、直接データ転送に関与しない記憶サブシステム１から記憶サブシステム３へ、"データ転送進捗問い合せ"コマンドが発行され、データセンタ３が応答して結果をデータセンタ１へ返す構成となっている。また図１０は、図１と図９を組み合せた構成である。記憶サブシステム３と５との間、記憶サブシステム２と５との間が、"データ進捗問い合せコマンド"の発行・応答の経路に該当する。

【００５５】上記の広域データストレージシステムの態様であれば、大規模な災害や、附近地に存在する２つのデータセンタに相次いで障害が発生した場合であっても、データセンタ３のホストへフェールオーバーすることで、災害直前のシステムが運用してきたデータを引き継いで処理でき、また、データの喪失を最小限度とすることができる。

【００５６】つまり、附近地にある２つのデータセンタが全滅する程度の災害が発生したときは、遠隔地に存在するデータセンタ３又は５（図１、図９、図１０）の記憶サブシステムを生かすことができる。ホストからの更新データの順序性が確保されつつ、非同期リモートコピーが行なわれているからである。但し、災害による未反映のデータは復旧できない。

【００５７】＜記憶サブシステムの構成について＞図１、図９及び図１０では、同期転送によるコピー及び非同期リモートコピーの組み合わせを示している。本来、リモートコピーは、１論理ボリュームと１論理ボリュームをデータ転送技術で結合したものである。本発明では、１個の論理ボリュームに対するデータ受領を、同期転送し、更に非同期転送して、附近地と遠隔地の双方にリモートコピー機能でデータ送信制御を行なっている。

【００５８】これらは記憶サブシステムの制御装置のマイクロコードで実現される機能である。ホストや別の記憶サブシステムからの更新データは、一旦、キャッシュ５（図２）に格納される。この時点では、当該データは、まだ記憶サブシステム内のハードディスクドライブにＲＡＩＤ制御により書き込まれていない。キャッシュ５内で当該データの転送制御情報を加え、別の記憶サブシステムへリモートコピー転送したり、複数の記憶サブシステムとのリモートコピー構成を同時に実現する制御を行う。同期転送と非同期転送による組合せを守ること

により、いつ災害が発生しても、各データセンタでは、データの更新順序を保った、データベースやジャーナルファイルシステムがリカバリ可能な論理ボリュームを常時、保持していることとなる。

【００５９】図２は、記憶サブシステムの概略構成を示す図である。

【００６０】制御装置１は、ホスト及びリモートコピーの接続先とデータの送受を行うチャネルアダプタ３、ディスク装置２内のハードディスクドライブ７をディスクインタフェース８（ディスクＩ／Ｆ８）を介して制御するディスクアダプタ９を有する。

【００６１】チャネルアダプタ３とディスクアダプタ９は、それぞれ、マイクロプロセッサを有し、データ転送バス１１・制御バス１１を介してキャッシュメモリ５と接続されている。尚、バス構成は一例であり、必要に応じてクロスバー構成としても良い。また、制御装置１を複数設けてクラスタ構成とし、複数の制御装置１を連絡する共通の第３のバスを追加しても良い。

【００６２】ホストとの間や、リモートコピーの接続先とデータ送受を行う際の格納元は、キャッシュ５である。制御情報、構成管理情報、転送状態／ビットマップは、制御メモリ６に格納されている。

【００６３】リモートコピーには送信及び受信の機能があり、本実施例ではホストからＩ／Ｏを受領するチャネルアダプタを分けて搭載している。ホストから受領したＩ／Ｏは、一旦、キャッシュ５へ格納される。リモートコピーの転送先情報や後述する状態管理／ビットマップは、制御データとして制御メモリ６に格納され、マイクロコードにより制御される。

【００６４】キャッシュに受領したデータは、ディスクアダプタ９によりハードディスクドライブ７へＲＡＩＤ制御で書き込まれる。これとは別の処理である、マイクロコードを用いた制御により、予め定義されたリモートコピー転送先への送信制御が行なわれる。

【００６５】例えば、ホストから受領したデータが、後続するリモートコピーの対象であり、非同期転送によるデータ送信を行うと定義されていた場合には、キャッシュ５の内部のデータに対して、データ受領順にシーケンス番号を付与する。これはデータ更新を示すＩＤ情報でもある。シーケンス番号を付与されたデータは、チャネルアダプタ３のリモートコピー送信機能により、当該シーケンス番号と共に送信される。

【００６６】別の実施例で、ホストから受領した更新ブロックを、複数の論理ボリュームと接続するリモートコピー制御が定義されていた場合には、キャッシュメモリ５の内部のデータは、同期転送用に加工されると同時に、非同期転送用にも加工され、シーケンス番号が付与されて、それぞれ、チャネルアダプタ３で附近地又は、遠隔地に向けて送信される。

【００６７】図２は本発明を実現する一例であり、本発明はハードウエア構成に依存しない。リモートコピー接続が記憶サブシステム間で実現可能であれば、マイクロプロセッサによる論理的なサポート、マイクロコード制御で実現できるためである。

【００６８】＜転送状態／ビットマップ＞図４は、転送状態／ビットマップ（適宜、ビットマップと略記する。）の一例を示したものである。これは、直接データ転送を行なっていない２つのデータセンタに設置された記憶サブシステムの内部に、災害・障害の復旧の際にペアを組むであろう相手（別のデータセンタに設置された記憶サブシステム）のデータ更新の進捗状況を知るために用意されたものである。例えば、図１ではデータセンタ２とデータセンタ３との間で、非常時のためにペアが組まれる。図９の広域データストレージシステムであれば、記憶サブシステム１と記憶サブシステム３との間で、図１０では、記憶サブシステム２と記憶サブシステム５、記憶サブシステム３と記憶サブシステム５との間で、それぞれ、非常時のために、ペアが組まれることとなる。

【００６９】転送状態／ビットマップは、ペア（対）となる論理ボリュームに対して必要であり、本発明では１個の論理ボリュームの実体に対し、２個以上の転送状態／ビットマップを持ち得る。各ビットマップは、ペアやペアとなる場合を想定した定義付けにより、相手の論理ボリュームとの差分管理を行うために使われる。ビットマップの中のブロックナンバは、論理ボリュームの更新を管理する最小単位であるブロックに対応させた番号である。

【００７０】ホストＩ／Ｏは、このブロックナンバと同一単位である必要はない。ホストＩ／Ｏの単位は、通常、最小で５１２バイトとされ上限も設けられているが可変である。一方、ビットマップは、５０ｋＢ弱の大きさ、又は７００ｋＢ程度の大きさのものもあるが、２０ｋＢから１０００ｋＢ程度まで種々の大きさがある。ホストＩ／Ｏの１ブロックに対して、必ずしも、１ビットマップが対応する訳ではない。

【００７１】ブロックナンバに対応するブロックの内容が更新されれば、差分管理は当該ブロックナンバ全体となり、同期（リシンク）を行うときに当該ブロックナンバのデータ全体が転送されることとなる。

【００７２】ビットマップは、ブロックナンバ毎に、当該論理ボリュームの更新された単位として、リモートコピーによるペアを再構築する際（再同期、リシンク）に当該更新されたブロックのみを転送する目的で、相手論理ボリュームに転送すべき"Ｕｐｄａｔｅ"情報を持つ。つまりＵｐｄａｔｅフラグがＯｎ（図４の実施例では１）、であれば転送対象であることを示す。通常のＵｐｄａｔｅは、ホストからのコマンド単位で為されることから、カウンタ値が０であることに基づき、Ｕｐｄａｔｅフラグを０とする。

【００７３】ビットマップは更に、同一ブロックナンバで、複数回の更新を記録するカウンタ値を持つ。カウンタ値は、更新が無ければ"０"、３回更新されれば"３"となる。ブロックナンバーで表されるデータブロックの大きさが、ホストから更新されるデータブロックより大きい場合には、このカウンタ値を使うことにより確実に相手論理ボリュームへ更新データのみを転送できることとなる。

【００７４】後述の"データ進捗問合せコマンド"の中に格納されたブロックナンバとカウンタ値と、問合先の記憶サブシステムのビットマップのブロックナンバとカウンタ値との比較を、データコピー監視機能（後述）で行う。この際に、ある記憶サブシステムが持つカウンタ値が、この記憶サブシステムに送付されて来た"データ進捗問合せコマンド"に記述されたカウンタ値と等しいか大きい場合に、所定の記憶サブシステムのビットマップのカウンタ値は１減算される処理を受ける。

【００７５】送付されて来た"データ進捗問合せコマンド"に記述されたカウンタ値未満である場合は、その記憶サブシステムのビットマップのカウンタ値は何ら処理を受けない。そして減算したかしないかを、"データ進捗問合せコマンド"に応答して返す。

【００７６】その記憶サブシステムのビットマップのカウンタ値が、送付されて来た"データ進捗問合せコマンド"に記述されたカウンタ値"以上"の場合には、データ更新の進捗は、正常なリモートコピー機能により既に、その記憶サブシステムにおいて格納済み、書き込み済みであることを意味する。また"未満"の場合には、データが未到着であることを意味している。

【００７７】図４のカウンタ値は有限であり、例えば、１バイト分をカウンタ値として割り当てた場合には、２５６回を超える管理はできない。この例では同一ブロックが２５６回を超えた更新を受けた場合には、最早、カウンタ値のＵｐを行わず、Ｕｐｄａｔｅフラグを恒久的に立ててしまう処理を行う。つまり図４でカウンタ値に"Ｏｖｅｒ　Ｆｌｏｗ"を意味する情報を格納する。

【００７８】このような恒久的な指定がなされると（図４、Ｏｖｅｒ　Ｆｌｏｗ）、ビットマップで特定される、恒久的指定のなされたブロックのＵｐｄａｔｅフラグの解除（０を入力すること）は、相手論理ボリュームへの転送が完了しコピーが確定したことを、このビットマップを有する記憶サブシステムが認識するときまで行なわない。

【００７９】カウンタ値を用いた更新管理を行う理由を次に補足説明する。

【００８０】例えば、５０ｋＢ程度のデータ量を有するトラックに対応させてビットマップの管理を行う場合に、この５０ｋＢのデータのうち、異なる３箇所が、異なる時刻において、それぞれ更新されたとする。トラックに対応させてビットマップ管理を行うのは、災害・障害後の復旧（再同期、リシンク）において扱う単位がトラック単位であるためである。

【００８１】カウンタ値による管理を行なわない場合には、Ｕｐｄａｔｅフラグのみ監視することとなるが、ある時刻でＵｐｄａｔｅフラグが１であることのみを確認しても、その後の時刻に２度目、３度目の更新があった場合には、２度目以降のデータ更新を見逃してしまう。新たにカウンタ値の概念を導入して、ホストからのコマンド単位で為される同一データブロック（ここではトラックの一部）の更新を微細に監視することで、かかる不都合を防ぐことができる。

【００８２】次に、図２の制御装置１の内部でマイクロコードにより実現される転送状態／ビットマップの機能について定義する。論理ボリュームはリモートコピーの対となる論理ボリュームとの間で下記の転送状態を有する。これらは同期転送又は非同期転送に依存しない。

【００８３】１）「正常ペア状態」とは、データの順序性を保証して、双方のボリューム間で、同一のデータを２重に保持している状態をいう。

【００８４】２）「転送抑止ビットマップ登録の状態」とは、データの更新をビットマップに登録する状態をいう。未だペアの相手へデータの転送は行なわれていない。

【００８５】３）「ビットマップ使用のコピー状態」とは、「転送抑止ビットマップ登録の状態」から「正常ペア状態」への移行期をいう。２重化のためのコピーの初期状態に当たる。

【００８６】４）「障害状態」とは、障害によりデータを転送できない状態をいう。ビットマップに登録される。

【００８７】５）「ペア無ビットマップ登録状態」とは、本発明固有の特殊な状態をいう。災害・障害前に、相互にデータ更新状態を監視し保持する必要から生じた状態である。

【００８８】６）「ペア無状態」とは、ビットマップは用意されているが、未だペアを組んでおらず、データ更新の情報が登録されていない状態をいう。

【００８９】「ペア無ビットマップ登録状態」が存在することが本発明の特徴となる。この状態を持つことなく、"転送抑止ビットマップ登録の状態"というサスペンド（Ｓｕｓｐｅｎｄ）状態で兼ねても良い。ここで、サスペンド状態とは、論理ボリュームへのデータの更新状態を、ビットマップでのみ管理し、リモートコピーによる転送制御を行なわない状態をいう。

【００９０】「ペア無ビットマップ登録状態」を持つのは、転送状態／ビットマップをペアで持つ必要からである（図３）。例えば、図１の広域データストレージシステムにおいては次の理由による。

【００９１】データセンタ３が保持するデータを監視するため、データセンタ２の記憶サブシステムの内部の論

理ボリュームに対応して設けられた転送状態／ビットマップに、データセンタ３のデータ更新状態を持つ必要があり、且つ、データセンタ２が保持するデータを監視するため、データセンタ３の記憶サブシステムの内部の論理ボリュームに対応して設けられた転送状態／ビットマップに、データセンタ２のデータ更新状態を持つ必要があるためである。

【００９２】図９の広域データストレージシステムにおいては、データセンタ２の障害発生に備えて、データセンタ１とデータセンタ３のリモートコピーの差分管理情報から、データセンタ１とデータセンタ３との間でペア構築を目的として、「ペア無ビットマップ登録状態」をデータセンタ１とデータセンタ３で持つ必要がある。この結果、記憶サブシステムやデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、障害回復後に更新部分のみの差分転送が可能となる。

【００９３】転送状態／ビットマップの機能は、上記の様な制御を実現するマイクロコード及びビットマップと関連する制御テーブルから成る。具体的機能は、例えば、図２のマイクロプロセッサ４のマイクロコードと制御メモリ６で行なわれるが、先に示した様にマイクロコードの制御により自由に実装できる。例えば、マイクロプロセッサ１０による実現も可能である。またマイクロプロセッサが１台のみの制御装置でも実現できる。

【００９４】＜広域データストレージシステムの運用＞図３は、図１の広域データストレージシステムが正常に運用されている場合の基本的な制御方法を説明するための概略図である。正常運転ではデータ進捗問合せコマンドを記憶サブシステム２から記憶サブシステム３へ送信する。例えば、記憶サブシステム１の障害の際、実際の差分データの転送に際しては、記憶サブシステム２と記憶サブシステム３との間で、転送状態／ビットマップの機能を使用し、両方の記憶サブシステムのビットマップについて、論理演算を行う。その結果に基づき、相当するデータブロックのみを記憶サブシステム２から記憶サブシステム３へ転送している。図８に、図１の広域データストレージシステムのデータセンタ１に障害・災害が発生した場合において、非同期リモートコピーを再開させる概略の手順を示す。

【００９５】図８において、正常な運用では、データセンタ１から附近地のデータセンタ２へ同期転送によりデータの二重化が図られる一方で、遠隔地のデータセンタ３へは非同期転送によりデータの更新順序を確保したコピーが行なわれている。そして、データセンタ２の記憶サブシステム２のスケジュールで、データ進捗問合せコマンドがデータセンタ３に対し発行され、データセンタ２と３とは管理情報を遣り取りして、データの差分管理を行なっている。

【００９６】データセンタ１に災害・障害が発生すると、データセンタ２の記憶サブシステムは、非同期転送により、差分データをデータセンタ３へ送付し、即時に、データセンタ２と遠隔地のデータセンタ３によるシステム運用を回復できる。

【００９７】図３において、転送状態／ビットマップは、１論理ボリューム当り２個持ち、それぞれが、これらのビットマップを用いた機能を有する。記憶サブシステム１は、記憶サブシステム２と記憶サブシステム３に対し、転送状態／ビットマップ＃１に対応する機能及びビットマップ＃２に対応する機能を持つ。

【００９８】記憶サブシステム２と記憶サブシステム３は、同期転送及び非同期転送の各々について転送状態／ビットマップ＃３及び＃６の機能をそれぞれ持つ。これら＃１と＃３、＃２と＃６のそれぞれの機能は、正常運転の際には、"正常ペア状態"を格納している。

【００９９】転送状態／ビットマップ＃４及び＃５の機能は、それぞれ、記憶サブシステム２及び記憶サブシステム３が持っている。この広域データストレージシステムが正常に運用されているときには、転送状態／ビットマップ＃４及び＃５の機能は、上述の"ペア無ビットマップ登録"状態を保持する。

【０１００】転送状態／ビットマップ＃４の機能は、記憶サブシステム３の論理ボリュームに対する差分管理を、転送状態／ビットマップ＃５の機能は、記憶サブシステム２の論理ボリュームに対する差分管理を、それぞれ行う。

【０１０１】図１０の拡張として、ホストからのＩ／Ｏを受領する、第１のデータセンタに設置された記憶サブシステムの制御装置１が、Ｎ台の同期転送のコピー先と、Ｍ台の非同期リモートコピーのコピー先を持つ構成では、その制御装置１は、Ｎ＋Ｍ個の転送状態／ビットマップの機能を有する。これに対応する、遠隔地又は附近地の記憶サブシステム（コピー先）も、転送状態／ビットマップを持つこととなる。この結果、制御装置１やデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、災害回復の際の更新部分のみの差分転送が可能となる。

【０１０２】＜データコピー監視機能＞次に、データコピー監視機能について説明する。この機能には、ビットマップの制御機能、リモートコピーのステータス管理機能、構成管理機能、データ進捗問合せコマンドの制御機能、リモートコピーのデータ転送指示機能等が含まれる。

【０１０３】図３の記憶サブシステム２の制御装置で、同期転送によるデータブロックを記憶サブシステム１から受領する。本データは記憶サブシステム２のキャッシュメモリに格納されディスクドライブで記憶される。この際、転送状態／ビットマップ＃４の機能により当該データブロックが登録される。図４のビットマップに登録

する。

【０１０４】次に当該ブロックナンバとカウンタ値を格納した"データ進捗問い合せ"コマンドを記憶サブシステム２から記憶サブシステム３に対して発行する。発行のタイミングは同期転送に基づいても良いし、記憶サブシステム２の独自のスケジュールで行なっても良い。

【０１０５】記憶サブシステム３の制御装置で、記憶サブシステム２からの"データ進捗問い合せ"コマンドを受領し、転送状態／ビットマップ＃４のブロックナンバとカウンタ値を切り出し、記憶サブシステム３の該当する転送状態／ビットマップ＃５のそれらと比較する。

【０１０６】その結果、転送状態／ビットマップ＃５のブロックナンバがＵｐｄａｔｅフラグ１（更新）を示し、かつ、カウンタ値が転送されて来たもの以上であれば、同期転送に係るデータと、非同期リモートコピーに係るデータとが一致しているので、転送状態／ビットマップ＃６の対応するブロックナンバから、カウンタ値を１減算する。

【０１０７】減算の結果、カウンタ値が"０"となった場合には、Ｕｐｄａｔｅフラグを"０"とする。カウンタ値が"Ｏｖｅｒ　Ｆｌｏｗ"である場合には、何も操作しない。

【０１０８】また、転送状態／ビットマップ＃５に登録されていたカウンタ値が、記憶サブシステム２からの問合せコマンドから抽出されたカウンタ値未満であったり、Ｕｐｄａｔｅフラグが"０"（Ｏｆｆ）で更新が示されなかった場合には、＃５への更新は行なわず、これをデータ進捗問合せコマンドの結果として記憶サブシステム２へ返す。

【０１０９】＃５の転送状態／ビットマップの機能が、＃６の転送状態／ビットマップのカウンタ値を減算するということは、記憶サブシステム１から同期転送により既に記憶サブシステム２へ到着したデータブロックが、記憶サブシステム１から記憶サブシステム３へ非同期転送により到着済であったことを意味している。

【０１１０】データコピー監視機能は、本応答結果を用いて記憶サブシステム２の転送状態／ビットマップ機能の制御を行なう。記憶サブシステム３で"データ進捗問い合せ"コマンドのブロックナンバとカウンタ値が既に登録されていた旨の応答を返す場合（減算できた場合）には、記憶サブシステム２の制御装置でも転送状態／ビットマップの機能でカウンタ値の減算、Ｕｐｄａｔｅフラグの操作を同様に行う。

【０１１１】当該コマンドの応答結果が、未登録であれば、記憶サブシステム１から記憶サブシステム３へのデータの非同期転送が未完であるとして、記憶サブシステム２の転送状態／ビットマップ＃４の機能は、自己のビットマップに更新状況を保持する。これは後に更新差分部分のみを再同期させる際の対象となる。

【０１１２】この時点で記憶サブシステム１が重大障害を持ち、記憶サブシステム２と記憶サブシステム３との間でリモートコピー構成を再構築（再同期、リシンク）しなければならない場合には、ビットマップを参照した結果、未転送のデータのみ、即ち、差分のデータブロックのみを、記憶サブシステム２から３へ転送すれば良い。その結果、差分データの転送だけで即時に"正常ペア"を構築できる。これを実現する機能を"データコピー監視機能"と呼ぶ。

【０１１３】＜正常な運用の際に、直接データ転送を行なわない記憶サブシステム間での差分管理方法、その１＞図９の広域データストレージシステムにおいて、記憶サブシステム２に障害が発生した場合に、記憶サブシステム１と記憶サブシステム３との間で非同期リモートコピーによるシステム運用の復旧を図るときを考える。

【０１１４】このために、ホストからデータ更新を受領した記憶サブシステム１の制御装置１（図２）は、記憶サブシステム２の制御装置１の論理ボリュームに同期転送のコピーによるデータ転送を行う際に次の処理を行なう。

【０１１５】転送するブッロクの位置情報を、記憶サブシステム１の制御装置１に存在するビットマップに、記憶サブシステム３の論理ボリュームの更新情報を格納する。このとき既に転送したブッロクが記憶サブシステム３において更新されていたときは、ビットマップのカウンタ値を１増加（インクリメント）する。

【０１１６】記憶サブシステム１の制御装置１は、記憶サブシステム２の制御装置１に対して同期転送が完了した後、記憶サブシステム３の制御装置１に対して、同期転送したデータブッロクが、記憶サブシステム２の制御装置１を経由して到着したか否かを問合せるため、記憶サブシステム１と記憶サブシステム３とを結ぶ通信線路を用いて、確認コマンドを発行する。

【０１１７】確認コマンドには、ホストから受領した更新データの記憶サブシステムにおけるデータブロックのブロックナンバとカウンタ値が含まれている。確認コマンドを受領した記憶サブシステム３の制御装置１は、記憶サブシステム２の制御装置１経由で既に存在するデータブロックが、確認コマンドで問合されたブロックと一致するか否かを判定する。

【０１１８】記憶サブシステム３の制御装置１は、記憶サブシステム２の制御装置１の論理ボリュームに対する転送状態／ビットマップの機能の他に、記憶サブシステム１の制御装置１の論理ボリュームに対する状態管理／ビットマップの機能を持つ。

【０１１９】記憶サブシステム３の制御装置１は、記憶サブシステム２の制御装置からデータを受領すると、記憶サブシステム１の制御装置１の状態を格納すべく、自己の持つ転送状態／ビットマップへ登録する。このビットマップでは、論理ボリューム内のアドレスに関わるブロック位置に対する更新情報を有し、さらに複数回の同

一ブロックに対する更新を管理するためにカウンタ値を有している。

【０１２０】記憶サブシステム３の制御装置１の転送状態／ビットマップに登録された結果は、記憶サブシステム１の制御装置１から発行された確認コマンドのブロックナンバ及びカウンタ値と比較される。比較の結果、一致又はカウンタ値が確認コマンドにあったカウンタ値以上の場合には、データの到着が正常に完了していると判断し、転送状態／ビットマップの機能を用いてビットマップのカウンタを１減算する。

【０１２１】他方、記憶サブシステム１の制御装置１は、記憶サブシステム３の制御装置１から返される結果が、記憶サブシステム３へデータブロックが記憶サブシステム２を経由して到着していることを示す場合には、上述の記憶サブシステム３の制御装置が為したように、転送状態／ビットマップの機能を用いてビットマップのカウンタを１減算する。

【０１２２】以上のように、ビットマップを監視・管理することで、記憶サブシステム２が災害等により重大障害を持ち、同期及び非同期転送によるデータ送受が行なえなくなった場合であっても、ホストがＩ／Ｏを発行する記憶サブシステム１と、記憶サブシステム２の内容を非同期リモートコピーにより格納した記憶サブシステム３との間で、非同期リモートコピーを構成することができる。

【０１２３】この際に記憶サブシステム１と３の、それぞれの制御装置の転送状態／ビットマップの機能により、論理ボリュームの全データをコピーすることなく、差分データのブロックのみを転送することにより、即時に、構築することができる。

【０１２４】＜正常な運用の際に、直接データ転送を行なわない記憶サブシステム間での差分管理方法、その２＞図１の広域データストレージシステムにおいて、ペアになっている論理ボリューム間、つまり、記憶サブシステム１と２並びに記憶サブシステム１と３、それぞれの間のデータ更新状態の管理のために、転送状態／ビットマップの機能が各論理ボリューム毎に用意される。

【０１２５】記憶サブシステム１の制御装置１で障害が発生し、同期転送のコピー及び非同期リモートコピーの双方が継続不可となった場合には、記憶サブシステム２と３のそれぞれの制御装置１の間で、先ず差分データをコピーし両者を一致させる。次いで、記憶サブシステム２と３の間で非同期リモートコピーを構成する。

【０１２６】ホストから更新すべきデータを受領した記憶サブシステム１の制御装置１は、記憶サブシステム２の制御装置１へ同期転送によりデータブロックを送出し、これを記憶サブシステム２の制御装置１が受領する。記憶サブシステム２の制御装置１は、受領したデータブロックの位置情報（ブロックナンバ）を、記憶サブシステム３の制御装置１の配下の論理ボリュームの管理情報との比較のために、自己が保持する転送状態／ビットマップに格納する。転送状態／ビットマップは、受領したデータブロックが更新された場合には、カウンタ値を１増加（インクリメント）する機能を備え、複数回のデータブロックの更新を記録できる。

【０１２７】記憶サブシステム２の制御装置１は、上記の転送状態／ビットマップへ所定の管理情報を登録した後、記憶サブシステム２の制御装置１と記憶サブシステム３の制御装置１との間を結ぶデータ転送経路を用いて、データブロックが記憶サブシステム３へ到着したか否かを問合せる確認コマンドを、記憶サブシステム３の制御装置１へ発行する。

【０１２８】確認コマンドは、記憶サブシステム２の制御装置１が、同期転送により記憶サブシステム１から受領したデータブロックの位置情報であるブロックナンバと、データブロックが何回更新されたかを示すカウンタ値を含む。

【０１２９】記憶サブシステム３の制御装置１は、記憶サブシステム１の制御装置１から、非同期リモートコピーの技術で受領したデータブロックの位置情報（ブロックナンバ）とカウンタ値を、記憶サブシステム２の制御装置１の配下の論理ボリュームの管理情報との比較のために、自己の制御装置１が持つ転送状態／ビットマップの機能を用いてビットマップに格納する。記憶サブシステム３の制御装置１は、ビットマップと確認コマンドの有する対応する値との比較を行う。

【０１３０】記憶サブシステム２から３へ問合せた確認コマンドが有するブロックナンバとカウンタ値と、記憶サブシステム３の制御装置１が持つ、記憶サブシステム２の制御装置１の配下の論理ボリュームの管理情報である、これらの値とを比較して、確認コマンドの値と同一又はカウンタ値が確認コマンドのカウンタ値より大きい場合には、転送状態／ビットマップの機能で、そのデータブロックのカウンタ値を１減算する。

【０１３１】減算した結果が０になる場合は、記憶サブシステム２と３との差分データはないことになるので、ビットマップの管理から削除する。上記の比較の結果が一致しない場合には、記憶サブシステム３の制御装置１は、ビットマップのカウンタ値を操作しない。

【０１３２】記憶サブシステム３の制御装置１は、記憶サブシステム２の制御装置１に、確認コマンドの応答である判定結果を返す。この結果を記憶サブシステム２の制御装置１が参照し、比較がカウンタ値を減算した場合には、既に記憶サブシステム２と３の間で、同一のデータブロックの更新が正常に終了していると断定する。

【０１３３】記憶サブシステム３に更新すべきデータブロックが届いていない場合には、記憶サブシステム２にのみ、更新に係るデータブロックが格納されていることになる。記憶サブシステム２の制御装置１は、自己の転送状態／ビットマップの機能でこれを記憶する。

【０１３４】記憶サブシステム２の制御装置１が、記憶サブシステム３の制御装置１から確認コマンドの応答を受領し、記憶サブシステム３に更新すべきデータブロックが未到着であった場合には、記憶サブシステム２の制御装置１が持つ、記憶サブシステム３の論理ボリュームの更新状態に対応する転送状態／ビットマップのカウンタ値は減算しない。このことは、そのビットマップは、更新に係るデータブロックが、記憶サブシステム２と３との間で差分であることを示す。

【０１３５】他方、データの到着完了を示した場合には、上記の転送状態／ビットマップの更新に係るデータブロックのカウンタ値を１減算する。カウンタ値が０のときは、記憶サブシステム２と３との間で、更新に係るデータブロックは同一であり不整合がないので、差分データのコピーの対象とはしない。

【０１３６】このように、正常運用の際に、直接データ転送を行っていない記憶サブシステムの制御装置同士が、災害・障害からの回復を想定して、論理ボリューム間の差分データ管理を行っているため、記憶サブシステム間で差分データのみをコピーし不一致をなくすことが高速に行なえる。

【０１３７】＜フェールオーバー後のシステムの運用＞図７に、図１の広域データストレージシステムが、フェールオーバーにより状態を遷移して図９の構成となった場合の運用について簡単に説明する。図３で記憶サブシステム１に、図９で記憶サブシステム２に、図１０で記憶サブシステム１、２又は４に、それぞれ重大な障害が起きた場合には、図７に示す様に、残存する２以上の記憶サブシステムの間で、リモートコピー構成の復帰を図ることとなる。

【０１３８】本発明によれば、図７の様に、直接データ転送に関与していなかった論理ボリューム間（記憶サブシステム１と記憶サブシステム３との間）で、差分データのみコピーすれば、即時に、リモートコピーのペアを生成でき、リモートコピーの運用再開が可能である。

【０１３９】本発明を実施しない場合には、図３の記憶サブシステム２と３の間、図９の記憶サブシステム１と３の間で、それぞれ、リモートコピー構成をつくるに際し、図３の構成では記憶サブシステム２から記憶サブシステム３に対し、図９の構成では、記憶サブシステム１から記憶サブシステム３に対し、それぞれ、記憶サブシステムが保持するデータのフルコピーを行なわなければならない。大規模のデータセンタでは、コピーに長時間を要し、リモートコピーの運用再開が遅くなる。長時間を要するコピー中に、再度、コピー元やデータ転送経路に障害・災害が発生すると、データは破壊され喪失することとなる。

【０１４０】図１１を用いて、図９の構成におけるデータコピー監視機能について簡単に説明する。

【０１４１】データ進捗問合せコマンドは記憶サブシステム１から記憶サブシステム３に対して発行される。データコピー監視機能は図１の場合と一部処理が異なる。記憶サブシステム１が、同期転送により記憶サブシステム２へホストから受領した更新データを転送した後、記憶サブシステム１から３に対し、上述した"データコピー監視機能"を作動させる。つまり、"データ進捗問い合せ"コマンドを発行し、記憶サブシステム１の持つ転送状態／ビットマップ＃１と、記憶サブシステム３の持つ転送状態／ビットマップ＃３で、それぞれのＵｐｄａｔｅフラグ、カウンタ値を登録し、所定の操作を行う。

【０１４２】記憶サブシステム１から３に、ホストから記憶サブシステム１が受領したデータ（トラック）と同じデータが、記憶サブシステム３に届いたか否か、問合せた結果、未着であれば、記憶サブシステム１の転送状態／ビットマップ＃１のビットマップは、そのまま保持する。結果が到着であれば、つまり、＃３のビットマップのブロックナンバ、カウンタ値が同一であれば、Ｕｐｄａｔｅフラグを削除し、＃１のビットマップを削除する。

【０１４３】＜再同期におけるその他の処理＞データコピー監視機能で検出した"データ進捗問い合せ"コマンドの応答結果に、エラーや不具合（タイムアウト）が生じたり、転送状態／ビットマップの機能に不具合が生じた場合には、障害・災害の際に行われるべき回復処理に関する差分管理を禁止する。

【０１４４】転送状態／ビットマップの機能において、ビットマップは有限なカウンタ値の格納領域を有している。この有限値を超えて（オバーフロー）、同一データブロックが更新された場合には、そのデータブロックは、その後２以上の記憶サブシステム間で冗長度が維持されていても、災害・障害発生後に再同期処理、差分コピー処理が行なわれる際に、必ず更新対象として扱う。

【０１４５】正常な運用において直接、データ転送を行なわない記憶サブシステム間で遣り取りされる問合せ（確認コマンド送出）に対し、所定時間、応答が無い場合は、タイムアウトであるとして再同期処理を禁止する。非同期リモートコピーによるペアの再構築処理や、差分データのみ転送する処理を行なわず、禁止する。ペアの相手のデータ更新状態を知ることができないため、そのままペアの再度構築処理を行なわしめることは妥当でないからである。

【０１４６】＜非同期転送におけるデータの整合性の管理＞例えば、ホストが接続する記憶サブシステム１と記憶サブシステム２とが、記憶サブシステム１から記憶サブシステム２にデータを複写する非同期転送で運用されているとする。この場合、もし、記憶サブシステム１におけるデータの書き込み順と、記憶サブシステム２におけるデータの書き込み順とが異なると、両記憶サブシステム１，２におけるデータの整合性が保証されなくなる。以下、このようなデータの不整合を回避するための

仕組みについて説明する。

【０１４７】まず、各記憶サブシステム１，２における記憶資源の記憶領域に所定サイズ（例えば、１６Ｋバイトごと）のブロックを区画して各ブロックに固有のブロック番号を割り当てる。そして、ホストからデータの書き込みがあったブロックについて、そのブロック番号とデータの書き込み順に付与したシーケンス番号との対応づけを制御メモリ６に管理する。例えば、図１２に示すように、ブロック番号が５６～５９のブロックにデータが書き込まれた場合には、図１３に示すデータ管理情報を制御メモリ６に作成する。

【０１４８】記憶サブシステム１から記憶サブシステム２への非同期転送に際しては、図１４の転送データフォーマットに示すように、転送するデータに前記データ管理情報を付帯させる。一方、これを受信した記憶サブシステム２では、図１５に示すように、データに付帯して送信されてきた前記データ管理情報を、制御メモリ６に管理する。ここで制御メモリ６に管理される前記データ管理情報、すなわち、シーケンス番号とブロックＩＤの組み合わせには、これに対応するデータのキャッシュメモリ上の位置情報も対応づけて記憶されている。記憶サブシステム２は、前記データ管理情報のシーケンス番号の順番にこれに対応するキャッシュメモリ上の前記位置情報に記憶されているデータを記憶資源に書き込んでいく。

【０１４９】以上のようにして、ホストが記憶サブシステム１の記憶資源に書き込んだ順番どおりに、記憶サブシステム２の記憶資源においてもデータが書き込まれ、両記憶サブシステム１，２におけるデータの整合が保証されることになる。

【０１５０】＜マルチホップ方式＞図１６（ａ）に示す広域データストレージシステムは、サイト１に設置された記憶サブシステム１と、サイト２に設置された記憶サブシステム２と、サイト３に設置された記憶サブシステム３とを備える。記憶サブシステム１には、この記憶サブシステム１を記憶手段として利用するホストが接続する。記憶サブシステム１と記憶サブシステム３との間も通信手段により接続される。

【０１５１】記憶サブシステム１と記憶サブシステム２とは、記憶サブシステム１から記憶サブシステム２にデータを複写する同期転送で運用されている。また、記憶サブシステム２と記憶サブシステム３とは、記憶サブシステム２から記憶サブシステム３にデータを複写する非同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチホップ方式」と称する。なお、マルチホップ方式における各記憶サブシステム間の通信を同期転送とするか、非同期転送とするかは任意に設定される。また、これら以外の転送方式であってもよい。

【０１５２】つぎに、図１６（ｂ）とともにマルチホップ方式によるデータ差分管理の詳細について説明する。

【０１５３】記憶サブシステム１は、ホストから書き込み対象データとその書き込み要求（Write I/O）とを受信すると（S121）、書き込み対象データを自身の論理ボリューム（第１の記憶資源）に書き込むとともに、書き込み処理を行った順にシーケンス番号を付与し、これと前記データが書き込まれた論理ボリューム（第１の記憶資源）上の位置（格納位置）を特定する書き込み位置情報とを対応づけて（所定のテーブルに）記憶する（S122）。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【０１５４】つぎに、記憶サブシステム１は、前記書込対象データを、これに付与された前記シーケンス番号とともに記憶サブシステム２に送信する（S123）。ここでこのような記憶サブシステム間で行われる、データとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには必要に応じて、前述したデータの書き込み位置情報が付帯される。

【０１５５】記憶サブシステム２は、記憶サブシステム１から送られてくる前記書き込み対象データとシーケンス番号とを受信して、これを自身の論理ボリューム（第２の記憶資源）に書き込む。記憶サブシステム２は、前記書き込み処理が完了すると、その完了通知を記憶サブシステム１に送信する。

【０１５６】記憶サブシステム２は、記憶サブシステム３に対し、適宜なタイミングで前記書き込み対象データと前記シーケンス番号とを送信する（S124）（なお、図１６（ｂ）では、時間差を表現するため、記憶サブシステム１から記憶サブシステム２に送信されるデータのシーケンス番号と、記憶サブシステム２から記憶サブシステム３に送信されるデータのシーケンス番号とを変えて表記している）。

【０１５７】つぎに、記憶サブシステム３は、前記データと前記シーケンス番号とを受信すると、前記書き込み対象データに対応して発行した前記シーケンス番号を、記憶サブシステム１に送信する（S125）。記憶サブシステム１は、記憶サブシステム３から送られてくるシーケンス番号を受信する。

【０１５８】ここで記憶サブシステム１は、受信したシーケンス番号と、自身が記憶しているシーケンス番号とこれに対応する書き込み位置情報との対応づけ（テーブル）を対照することで、記憶サブシステム３の論理ボリューム（第３の記憶資源）に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム３から受領した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することにより行われる（S126）。

【０１５９】以上のようにしてマルチホップ方式におけ

る通常運用が行われる。

【０１６０】つぎに、災害等により記憶サブシステム２が停止した場合の回復処理について説明する。

【０１６１】図１７（ａ）に示すように、記憶サブシステム１は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム２の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが途切れるなどして、記憶サブシステム１が記憶サブシステム２の障害発生を検知した場合に、記憶サブシステム１と記憶サブシステム３の間を、差分データのみを複写することによってその内容を一致させ、その後記憶サブシステム１と記憶サブシステム３の間を、非同期転送での臨時運用へ移行させる処理について、図１７（ｂ）とともに説明する。

【０１６２】記憶サブシステム１は、記憶サブシステム２の障害発生を検知した場合（S131）、まず、制御メモリ６上に、自身の論理ボリューム（第１の記憶資源）の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム３において未反映の前記差分データについての前記シーケンス番号と前記書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップに対応する位置のビットをオンにする（S132）。

【０１６３】つぎに、記憶サブシステム１の論理ボリュームの、前記ビットマップ上のオンになっている位置に格納されている差分データを、記憶サブシステム１から記憶サブシステム３の対応する格納位置に複写する（S133）。そして、この複写完了後、記憶サブシステム１から非同期転送により差分データが複写される形態で、臨時運用が開始される（S134）。

【０１６４】ここでこの臨時運用への切り替えに際しては、記憶サブシステム２に障害が発生した場合でも、記憶サブシステム１のデータを記憶サブシステム３に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば、記憶サブシステム１と記憶サブシステム３との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを容易に同期させることができる。

【０１６５】つぎに、記憶サブシステム２が復旧し、臨時運用から通常運用に切り替える際の一連の処理について説明する。

【０１６６】まず、記憶サブシステム１は、自身の論理ボリューム（第１の記憶資源）に記憶している全てのデータを記憶サブシステム２の論理ボリューム（第２の記憶資源）に複写した後、記憶サブシステム１から記憶サブシステム２にデータを複写する同期転送での運用を開始する。すなわち、記憶サブシステム１は、ホストからの指示により自身の論理ボリューム（第１の記憶資源）にデータ書き込みを行った場合、書き込んだデータとシーケンス番号とを記憶サブシステム２に送信する。

【０１６７】記憶サブシステム２は、記憶サブシステム１から送られてくる前記書き込んだデータとシーケンス番号とを受信して、これを自身の論理ボリューム（第２の記憶資源）に書き込む。記憶サブシステム２は、前記書き込み処理が完了すると、自身の論理ボリューム（第２の記憶資源）に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号とを対応づけて（所定のテーブルに）記憶する。この段階のデータ転送状態を図１８に示す。

【０１６８】つぎに、記憶サブシステム３は、記憶サブシステム１から送られてくる前記データと前記シーケンス番号とを受信して、前記データを自身の論理ボリューム（第３の記憶資源）に記憶するとともに（図１８）前記対応づけにおける前記シーケンス番号を記憶サブシステム２に送信する（図示せず）。

【０１６９】記憶サブシステム２は、記憶サブシステム３から送られてくるシーケンス番号を受信する。ここで記憶サブシステム２は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号と、これに対応する書き込み位置情報とを対照することで、記憶サブシステム３の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【０１７０】つぎに、臨時運用において記憶サブシステム１から記憶サブシステム３に複写する非同期転送の運用を停止する。この停止後、記憶サブシステム２は、自身の制御メモリ上に、自身の論理ボリューム（第２の記憶資源）の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム３において未反映の前記差分データについてのシーケンス番号と書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【０１７１】つぎに、記憶サブシステム２は、前記ビットマップにより把握した、記憶サブシステム３の論理ボリューム（第３の記憶資源）において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム３に送信する。

【０１７２】記憶サブシステム３は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを、自身の論理ボリューム（第３の記憶資源）の、前記書き込み位置情報により指定される該当データの格納位置に記憶する。これにより、記憶サブシステム２の論理ボリューム（第２の記憶資源）の内容と、記憶サブシステム３の論理ボリューム（第３の記憶資源）の内容との同期が取れることになる。以上の処理終了後、記憶サブシステム２と記憶サブシステム３との間の非同期転送による運用が開始され、図１９に示す通常状態での運用が再開する。

【０１７３】以上のようにして臨時運用から通常運用への切り替えが完了する。

【０１７４】＜マルチコピー方式＞図２０に示す広域データストレージシステムは、サイト１に設置された記憶サブシステム１と、サイト２に設置された記憶サブシステム２と、サイト３に設置された記憶サブシステム３とを備える。記憶サブシステム２にはこの記憶サブシステム２を記憶手段として利用するホストが接続する。なお、記憶サブシステム１と記憶サブシステム３との間も通信手段により接続される。

【０１７５】記憶サブシステム１と記憶サブシステム２とは、記憶サブシステム２から記憶サブシステム１にデータを複写する同期転送で運用されている。また、記憶サブシステム２と記憶サブシステム３とは、記憶サブシステム２から記憶サブシステム３にデータを複写する非同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチコピー方式」と称する。なお、マルチコピー方式において、各記憶サブシステム間の通信を同期転送とするか、非同期転送とするかは前記の形態に限られず、任意に設定される。また、同期転送や非同期転送以外の転送方式であってもよい。

【０１７６】つぎに、図２０とともにこの実施例のデータ差分管理方式について説明する。記憶サブシステム２は、ホストから書き込み対象データとその書き込み要求（Write I/O）とを受信すると（S161）、書き込み対象データを自身の論理ボリューム（第２の記憶資源）に書き込む。また、記憶サブシステム２は、書き込まれたデータと、書き込み処理を行った順に付与したシーケンス番号とを、記憶サブシステム１に送信する（S162）。そして同時に、前記書き込まれたデータと前記付与したシーケンス番号とを、記憶サブシステム３に送信する（S164）。なお、前述のマルチホップ方式の場合と同様に、このような記憶サブシステム間で行われるデータとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには、必要に応じて前述したデータの書き込み位置情報が付帯される。

【０１７７】つぎに、記憶サブシステム１は、記憶サブシステム２から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム（第１の記憶資源）に書き込む。その際、前記シーケンス番号と、これと前記データが書き込まれた論理ボリューム（第１の記憶資源）上の位置（格納位置）を特定する書き込み位置情報とを対応づけて（所定のテーブルに）記憶する（S163）。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【０１７８】つぎに、記憶サブシステム３は、記憶サブシステム２から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム（第３の記憶資源）に書き込む。書き込みが完了すると、記憶サブシステム３は記憶サブシステム１に対し、前記書込み対象データとこれと対になっていた前記シーケンス番号とを記憶サブシステム１に送信する（S165）。記憶サブシステム１は、記憶サブシステム３から送られてくるシーケンス番号を受信する。

【０１７９】ここで記憶サブシステム１は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号とこれに対応する書き込み位置情報との対応づけを対照することで、記憶サブシステム３の論理ボリューム（第３の記憶資源）に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム３から受領した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することで行われる（S166）。

【０１８０】以上のようにしてマルチコピー方式における通常運用が行われる。

【０１８１】つぎに、災害等により記憶サブシステム２が停止した場合の回復処理について説明する。

【０１８２】図２１（ａ）に示すように、記憶サブシステム１は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム２の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが途切れるなどして、記憶サブシステム１が記憶サブシステム２の障害発生を検知した場合に、記憶サブシステム２に接続するホストに代えて、記憶サブシステム１と記憶サブシステム３の間を、差分データのみを複写することによってその内容を一致させ、その後記憶サブシステム１と記憶サブシステム３の間を、非同期転送での臨時運用へ移行させる処理について、図２１（ｂ）とともに説明する。

【０１８３】記憶サブシステム１は記憶サブシステム２の障害発生を検知した場合（S171）、例えば、オペレータの操作により、記憶サブシステム２に接続していたホストの業務の運用が、記憶サブシステム１に接続する副ホストに引き継がれる。

【０１８４】つぎに、記憶サブシステム１は、制御メモリ６上に、自身の論理ボリューム（第１の記憶資源）の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム３において未反映の前記差分データについてのシーケンス番号とデータ更新位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする（S172）。

【０１８５】つぎに、記憶サブシステム１の論理ボリュームの、前記ビットマップ上のオンになっている位置に対応する位置に格納されている差分データを、記憶サブシステム１から記憶サブシステム３に複写する（S173）。そして、複写完了後、記憶サブシステム１から同期転送によりデータが複写される形態で、臨時運用が開

始される（S174）。

【０１８６】ここでこの臨時運用への切り替えに際しては、記憶サブシステム２に障害が発生した場合でも、記憶サブシステム１のデータを記憶サブシステム３に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば記憶サブシステム１と記憶サブシステム３との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを簡単に同期させることができる。

【０１８７】つぎに、記憶サブシステム２が復旧し、臨時運用から通常運用に切り替える際の一連の処理について説明する。

【０１８８】まず、記憶サブシステム１は、自身の論理ボリューム（第１の記憶資源）に記憶している全てのデータを記憶サブシステム２の論理ボリューム（第２の記憶資源）に複写した後、記憶サブシステム１から記憶サブシステム２にデータを複写する同期転送での運用を開始する。なお、このとき記憶サブシステム１と記憶サブシステム３間での非同期転送も継続して行われる。

【０１８９】記憶サブシステム１は、ホストから書き込まれたデータと、書き込み処理を行った順に付与したシーケンス番号とを、記憶サブシステム２に送信する。そして同時に、前記書き込まれたデータと前記付与したシーケンス番号とを、記憶サブシステム３にも送信する。

【０１９０】記憶サブシステム２は、自身の論理ボリューム（第２の記憶資源）に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶する（位置情報管理テーブル作成）。この段階での運用状態を図２２に示す。

【０１９１】記憶サブシステム３は、記憶サブシステム１から送られてくる前期データと前記シーケンス番号とを受信して、前記データを自身の論理ボリューム（第３の記憶資源）に記憶するとともに前記対応づけにおける前記シーケンス番号を記憶サブシステム２に送信する。

【０１９２】記憶サブシステム２は、記憶サブシステム３から送られてくるシーケンス番号を受信する。ここで記憶サブシステム２は、前記受信したシーケンス番号と、自身が記憶している前記対応づけを対照することで、記憶サブシステム３の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【０１９３】つぎに、臨時運用において記憶サブシステム１から記憶サブシステム３に複写する非同期転送の運用を停止する。この停止後、記憶サブシステム２は、自身の制御メモリ上に、自身の論理ボリューム（第２の記憶資源）の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム３において未反映の前記差分データについてのシーケンス番号と書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【０１９４】つぎに、記憶サブシステム２は、前記ビットマップにより把握した、記憶サブシステム３の論理ボリューム（第３の記憶資源）において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム３に送信する。

【０１９５】記憶サブシステム３は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて自身の論理ボリューム（第３の記憶資源）に記憶する。これにより、記憶サブシステム２の論理ボリューム（第２の記憶資源）の内容と、記憶サブシステム３の論理ボリューム（第３の記憶資源）の内容との同期が取れることになる。それから記憶サブシステム２から記憶サブシステム３への非同期転送が開始される。この段階での運用状態を図２３に示す。

【０１９６】ここで記憶サブシステム１に接続するホストの記憶サブシステム１へのデータ書き込み処理が完了しており、記憶サブシステム１と記憶サブシステム２の同期が取れている時に、記憶サブシステム１から記憶サブシステム２に対して行っていたデータの複写を、記憶サブシステム２から記憶サブシステム１に対して行うように切り替える。すなわち、同期が取れている状態で切り替えを行うことで、差分データを複写する等の作業が必要でなくなる。

【０１９７】つぎに、記憶サブシステム１に接続するホストにより運用されている業務を、記憶サブシステム２に接続するホストに引き継ぐ。そして、記憶サブシステム２から記憶サブシステム３にデータを複写する同期転送による運用を開始することで、図２４に示す通常状態での運用が再開することになる。

【０１９８】以上のようにして臨時運用から通常運用への切り替えが完了する。

【０１９９】＜他の障害復旧方式＞つぎに、障害復旧方式のバリエーションについて説明する。

【０２００】図２５に示すマルチホップ方式において、記憶サブシステム１がダウンした場合（ａ）には、記憶サブシステム２に副ホストを接続し、この副ホストにより記憶サブシステム１に接続するホストの業務を引き継ぐ。なお、記憶サブシステム２と記憶サブシステム３の間では、非同期転送での運用が行われている（ｂ）。

【０２０１】記憶サブシステム１が復旧した場合には、まず、記憶サブシステム２の全データを記憶サブシステム１に複写し、副ホストの業務を記憶サブシステム１に接続するホストに引き継ぐ。そして、前記の要領で、記憶サブシステム１と記憶サブシステム２との間のデータ転送方向を逆向きにすることにより、通常運用を再開する（ｃ）。

【０２０２】図２６に示すマルチホップ方式において、

記憶サブシステム３に障害が発生した場合（ａ）には、記憶サブシステム３の復旧後、記憶サブシステム２から記憶サブシステム３に全データを複写して記憶サブシステム３のデータを記憶サブシステム２と同期させ、記憶サブシステム１から記憶サブシステム２にデータを複写する同期転送および記憶サブシステム２から記憶サブシステム３にデータを複写する非同期転送による通常運用を再開する（ｂ）。

【０２０３】図２７に示すマルチコピー方式において、記憶サブシステム１に障害が発生した場合（ａ）には、記憶サブシステム１の復旧後、記憶サブシステム２から記憶サブシステム１に全データを複写して記憶サブシステム１のデータを記憶サブシステム２と同期させ、記憶サブシステム２から記憶サブシステム１にデータを複写する同期転送および記憶サブシステム２から記憶サブシステム３にデータを複写する非同期転送による通常運用を再開する（ｂ）。

【０２０４】図２８に示すマルチコピー方式において、記憶サブシステム３に障害が発生した場合には、記憶サブシステム３の復旧後、記憶サブシステム２から記憶サブシステム３に全データを複写して記憶サブシステム３のデータを記憶サブシステム２と同期させ、記憶サブシステム２から記憶サブシステム１にデータを複写する同期転送および記憶サブシステム２から記憶サブシステム３にデータを複写する非同期転送による通常運用を再開する。

【０２０５】＜複写元・複写先、書き込み位置情報の管理＞記憶サブシステム間でデータを転送する場合、データの転送元や転送先の設定や、その転送が同期・非同期いずれの方式で行われるかといった設定は、オペレータが各記憶サブシステムを操作して設定する場合（なお、この場合には、例えば、ある記憶サブシステムが障害を起して使えなくなった場合に、どの記憶サブシステムが次のデータの転送元になり、どの記憶サブシステムが次の転送先になるのかということを、システムの構成時に予め登録しておく）、記憶サブシステムに付帯するシステムが自動的に行ようにしている場合など、システムの構成に応じて様々な形態で行われる。

【０２０６】また、シーケンス番号と書き込み位置情報の対応づけの管理は、例えば、オペレータが、転送元や転送先を記憶サブシステムに登録する操作を開始する契機で行う。

【０２０７】＜記憶サブシステムの選択方式＞図２９に示す広域データストレージシステムは、記憶サブシステム１とこれに接続するホスト１ｈ、記憶サブシステム１からデータが非同期転送される記憶サブシステム２および記憶サブシステム３を備えている。ホスト１ｈもしくは記憶サブシステム１に障害が発生した場合、迅速に記憶サブシステム２もしくは記憶サブシステム３のどちらか一方を主たる記憶サブシステムとして選択し、また、信頼性・保全性確保のため、これら２つの記憶サブシステム２および３においてデータを２重化管理する。以下、ホスト１ｈもしくは記憶サブシステム１に障害が発生した場合に行われる処理について説明する。

【０２０８】記憶サブシステム２は、例えば、記憶サブシステム１から送信されてくるデータの有無や、記憶サブシステム１からあらかじめ設定された時間等に送られてくるハートビートメッセージの監視により、ホスト１ｈや記憶サブシステムに障害が発生したことを検知する。

【０２０９】障害を検知した場合、記憶サブシステム２は、迅速に主たる記憶サブシステムを決定し、副ホスト２もしくは副ホスト３による臨時運用に切り替える。主たる記憶サブシステムの選択はつぎのようにして行われる。まず、障害を検知した記憶サブシステム２は、記憶サブシステム３に、前述したシーケンス番号のうち最新のシーケンス番号の送信を要求するメッセージを送信する。記憶サブシステム３は、前記メッセージを受信すると、自身が記憶している最新のシーケンス番号を記憶サブシステム２に送信する。

【０２１０】記憶サブシステム２は、記憶サブシステム３から送られてきたシーケンス番号と、自身が記憶している最新のシーケンス番号とを比較して、より最新のシーケンス番号を受信している記憶サブシステムを、主たる記憶サブシステムとして選出し、選出した記憶サブシステムの識別子を選出候補として記憶するとともに、前記識別子を記憶サブシステム３に送信する。記憶サブシステム３は、送信されてきた前記識別子を受信し、これによりどの記憶サブシステムが主たるサブシステムとして選出されたのかを認知する。

【０２１１】なお、以上の選出処理において、記憶サブシステム間の通信方式の性質などの諸事情により、記憶サブシステム２もしくは記憶サブシステム３が記憶しているシーケンス番号に抜けが存在することがある。そこで、このような場合には、連続しているシーケンス番号のうちで最新のものを、前記の比較に用いる。

【０２１２】主たる記憶サブシステムが選出されると、つぎに、記憶サブシステム２と記憶サブシステム３とによりデータの二重化管理を行うため、両者が記憶しているデータの内容を一致させる。これは、記憶サブシステム間で全データの複写や差分データの複写により行われる。記憶サブシステム間でデータが一致すると、主たる記憶サブシステムとして選出された記憶サブシステムは、自身に接続している副ホストに、自身が主たる記憶サブシステムとなる旨を送信する。副ホストはこれを受信して代行運用を開始する。また、記憶サブシステム２と記憶サブシステム３との間で、同期転送もしくは非同期転送によるデータの二重化管理が開始される。

【０２１３】なお、以上の説明では、記憶サブシステム２が記憶サブシステム３から最新のシーケンス番号を取

得して、主たる記憶サブシステムを選出するようにしているが、この処理は記憶サブシステム３が行ってもよい。

【０２１４】また、記憶サブシステム１乃至記憶サブシステム３の３台構成の記憶サブシステムにおいて、記憶サブシステム１の障害発生時に代行して運用される他の記憶サブシステムを選出する仕組みを一例として説明したが、前述の仕組みは、４台以上の記憶サブシステムで構成される広域データストレージシステムにも適用することができる。

【０２１５】＜キャッシュメモリ上のデータの管理＞ホストが接続する一次の記憶サブシステムに、この一次の記憶サブシステムのデータのリモートコピー先である１以上の二次の記憶サブシステムが接続する系における、一次の記憶サブシステムのキャッシュメモリ上のデータの管理に関する実施例について説明する。

【０２１６】前記の系において、一次の記憶サブシステムから二次の記憶サブシステムに複写（リモートコピー）する必要の無いデータについては、一次の記憶サブシステムの記憶資源にデータを書き込んだ後は、そのデータを当該記憶サブシステムのキャッシュメモリ上から消去されてもよいが、二次の記憶サブシステムに複写する場合には、少なくともそのデータを二次の記憶サブシステムに送信するまではキャッシュメモリ上に残しておく必要がある。また、転送先となる二次の記憶サブシステムが複数存在する場合には、通信手段の違いや運用上の差異などにより、通常、二次の記憶サブシステムについての転送が同時に行われるわけではないので、このような場合には、全ての二次の記憶サブシステムに対する転送が終了するまで、データをキャッシュメモリ上に残しておく仕組みが必要である。

【０２１７】そこで、一次の記憶サブシステムにおいて、キャッシュ上に置かれているデータについて、一次の記憶サブシステムに接続する二次の各記憶サブシステムについての転送が完了しているかどうかを管理するようにする。具体的には、例えば、図３０に示すように、キャッシュメモリ上に区画された記憶ブロック（＃１，～，＃ｎ）ごとに、それぞれの記憶ブロックに格納されているデータについて、二次の各記憶サブシステムへの転送が完了しているかどうかを示すテーブルを、一次の記憶サブシステムにおいて管理するようにする。

【０２１８】なお、このテーブルにおいて、ビット「０」は転送が完了していることを示し、ビット「１」は転送が完了していないことを示す。ホストからのデータが一次の記憶サブシステムに書き込まれた時に、データが書き込まれた記憶ブロックの転送先となっている二次の記憶サブシステムに対応するビットに「１」がセットされる。ある記憶ブロックの「１」がセットされているビットのうち、データの転送が完了した二次の記憶サブシステムについてのビットは、転送完了後に「０」となる。

【０２１９】そして、全ての二次の記憶サブシステムについてのビットが「０」となった記憶ブロックに格納されているデータについては、キャッシュメモリ上から消去してもよいということになる。

【０２２０】

【発明の効果】図１、図９及び図１０で示した、３つ以上のサイトを有する広域データストレージシステムにおいて、いずれかのサイトに、いつ災害・障害が発生しても、巨視的に見て常時、データの順序性を保証した論理ボリュームを残すことができる。

【０２２１】直接データ転送に関与していなかった論理ボリューム間、例えば、図７の記憶サブシステム１と記憶サブシステム３との間で、差分データのみコピーすれば、即時に、非同期リモートコピーのペアを生成でき、広域データストレージシステムの運用再開が、即時に、可能となる効果がある。

【０２２２】本発明では、記憶サブシステムの内部にリモートコピーを実施するための冗長な論理ボリュームを必要としないため、記憶サブシステムのメモリー資源の使用効率が上がり、記憶サブシステムのコストパフォーマンスが向上する効果がある。

【図面の簡単な説明】

【図１】本発明に係る広域データストレージシステムの全体構成の一例を示した説明図である。

【図２】記憶サブシステムの一例を示した概念図である。

【図３】図１の構成において、データコピー監視機能を説明するための概念図である。

【図４】本発明を実現するための転送状態／ビットマップの一例を示した図である。

【図５】一般的な同期転送によるコピーの制御の概略を説明するための図である。

【図６】非同期リモートコピーの制御の概略を説明するための図である。

【図７】図９の全体構成において、データセンタ２に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図８】図１の全体構成において、データセンタ１に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図９】本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図１０】データセンタを４拠点以上設置した場合の、本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図１１】図９の全体構成において、データコピー監視機能を説明するための概念図である。

【図１２】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶資源

のデータを管理する単位であるブロックの概念を示す図である。

【図１３】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、データ管理情報の概念を示す図である。

【図１４】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、転送データのフォーマットの概念を示す図である。

【図１５】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶サブシステム２において管理されるデータ管理情報の概念を示す図である。

【図１６】（ａ）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（ｂ）は（ａ）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図１７】（ａ）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（ｂ）は（ａ）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図１８】マルチホップ方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図１９】マルチホップ方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図２０】（ａ）はマルチコピー方式の広域データストレージシステムの概念を示す図であり、（ｂ）は（ａ）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図２１】（ａ）はマルチコピー方式の広域データストレージシステムの概念を示す図であり、（ｂ）は（ａ）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図２２】マルチコピー方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図２３】マルチコピー方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図２４】マルチコピー方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図２５】（ａ）～（ｃ）は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図２６】（ａ）、（ｂ）は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図２７】（ａ）、（ｂ）は、マルチコピー方式における障害復旧方式の他のバリエーションを説明する図である。

【図２８】（ａ）、（ｂ）は、マルチコピー方式における障害復旧方式の他のバリエーションを説明する図である。

【図２９】障害発生時において、本番業務を代行させる記憶サブシステムの選択方式を説明する、広域データストレージシステムの概念図である。

【図３０】本発明の一実施例による、キャッシュメモリ上のデータの管理方法における、二次の各記憶サブシステムへのデータの転送状態を管理するテーブルを示す図である。

【符号の説明】

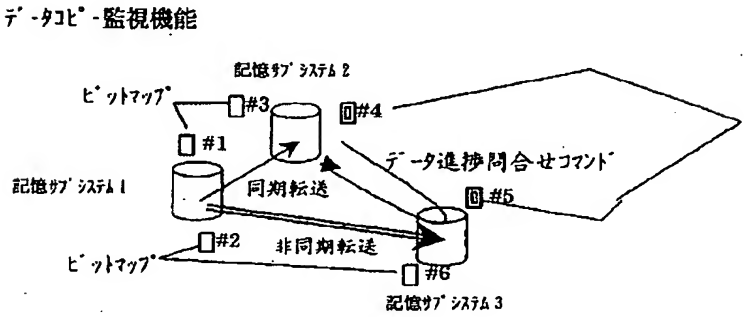
１　記憶サブシステムの制御装置

５　キャッシュメモリ

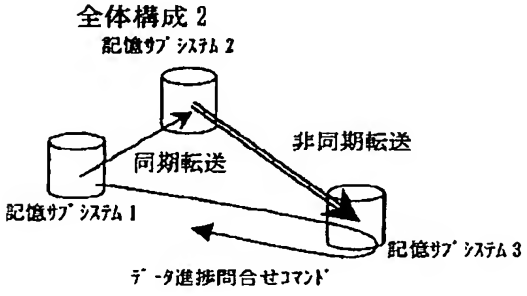
６　制御メモリ

＃１～＃６　転送状態／ビットマップ。

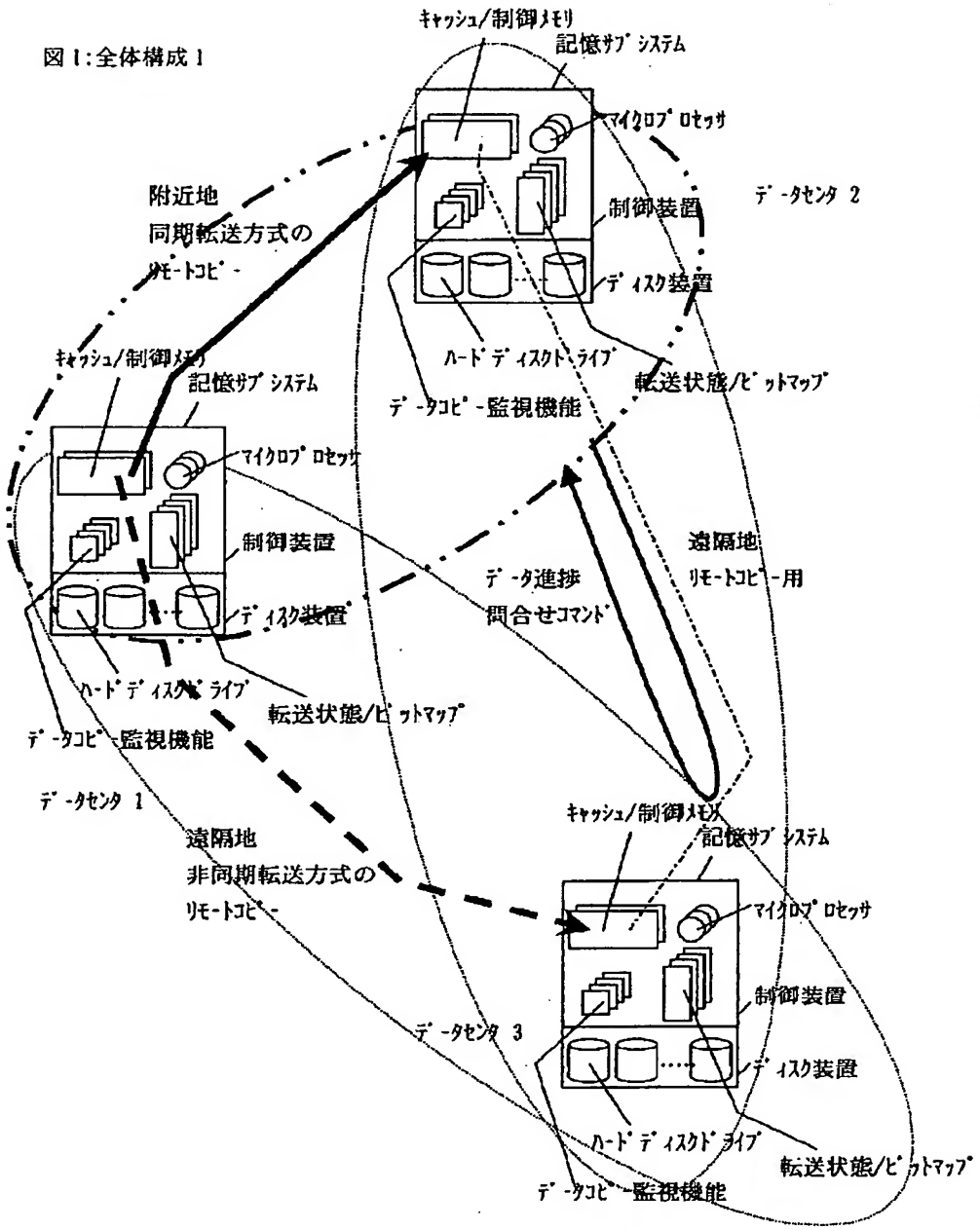
【図３】



【図９】



【図１】

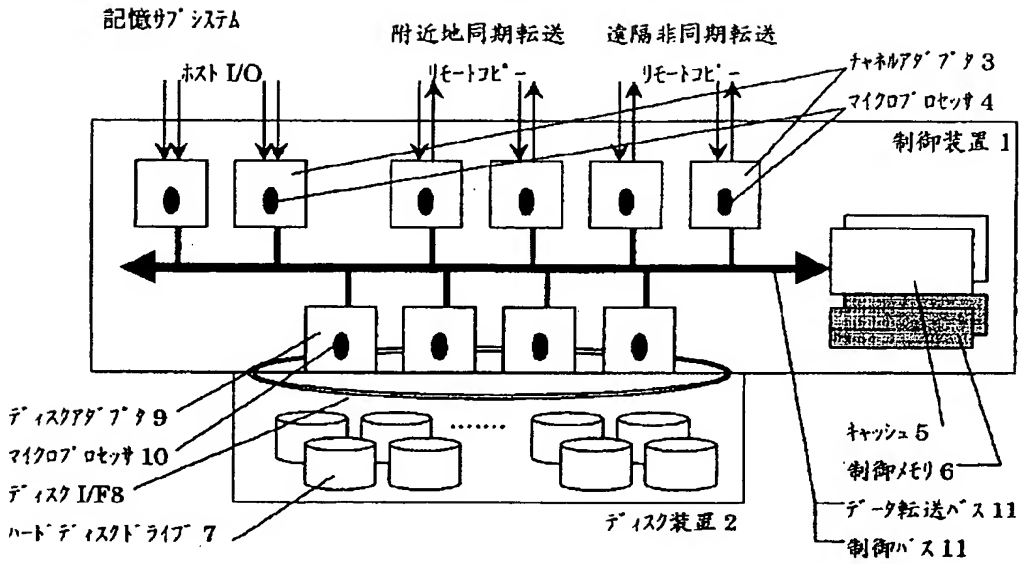


【図４】

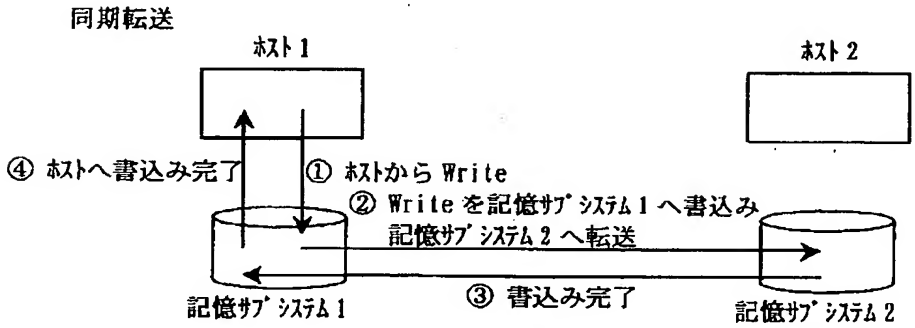
転送状態／ﾋﾞｯﾄﾏｯﾌﾟ

ﾌﾞﾛｯｸﾅﾝﾊﾞｰ	Update ﾌﾗｸﾞ	ｶｳﾝﾀ値	-----
----	---	---	---
3030	1	3	---
3031	0	0	---
3032	1	2	---
---	---	---	---
4032	1	Over Flow	---
---	---	---	---

【図２】



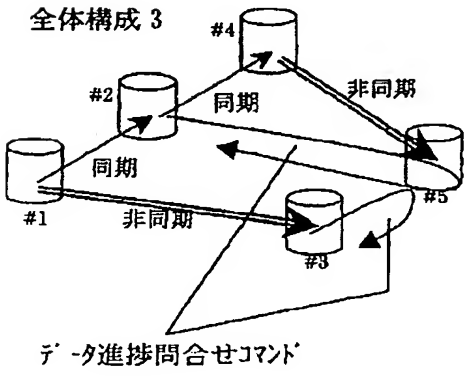
【図５】



【図１３】

シーケンス番号	0
ブロック番号	５６
シーケンス番号	1
ブロック番号	５７
シーケンス番号	2
ブロック番号	５８
シーケンス番号	3
ブロック番号	５９

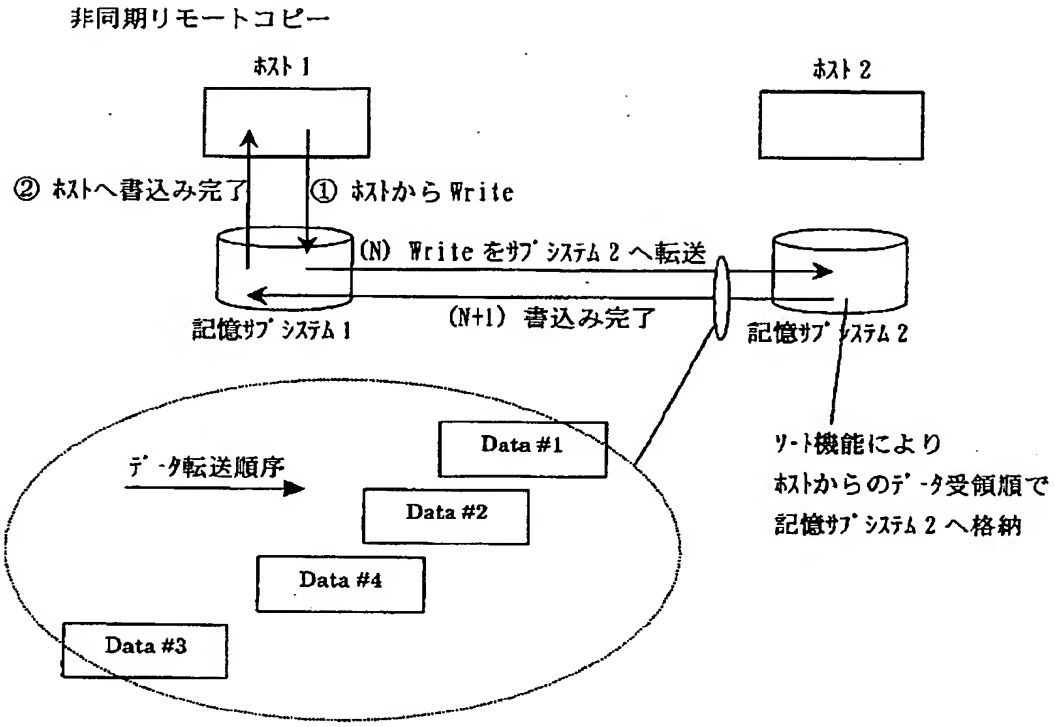
【図１０】



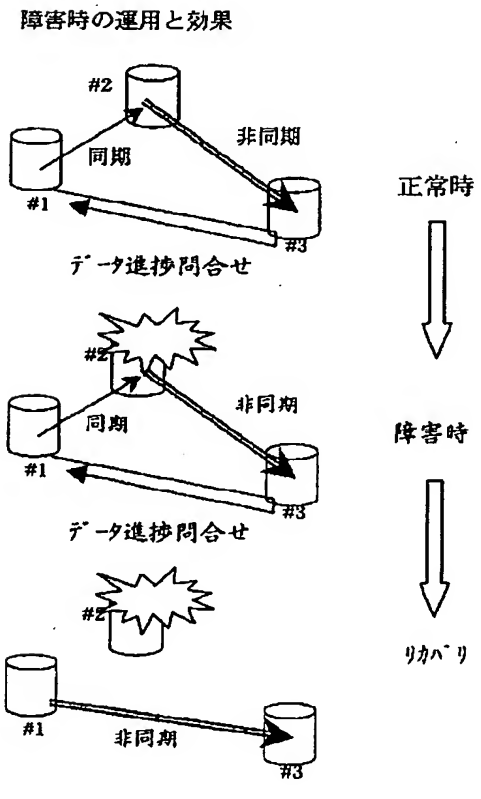
【図１５】

シーケンス番号	0
ブロック番号	５６
キャッシュ番号	１２２
シーケンス番号	3
ブロック番号	５９
キャッシュ番号	2
シーケンス番号	1
ブロック番号	５７
キャッシュ番号	１６

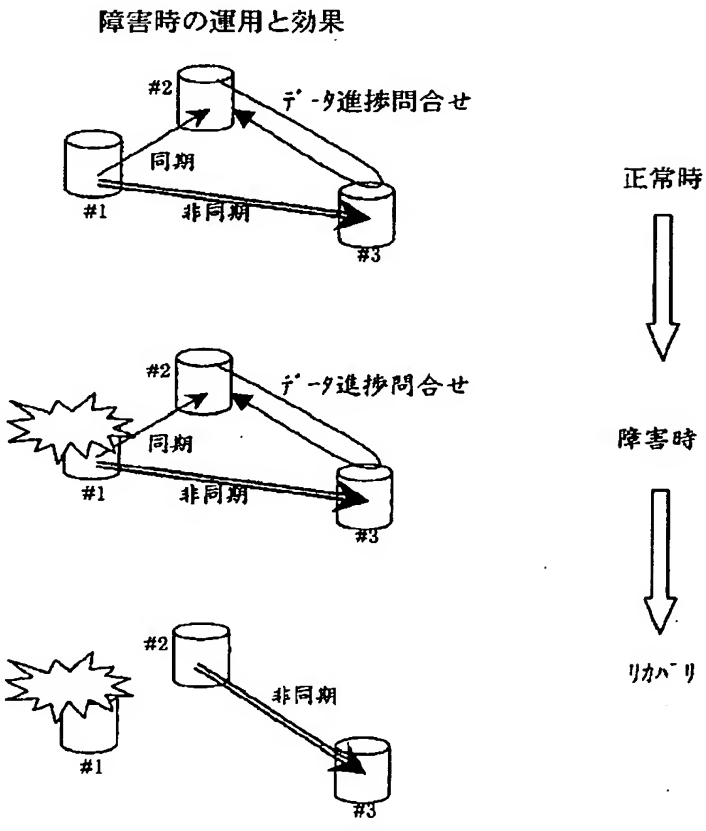
【図６】


非同期リモートコピー
ホスト１
ホスト２
② ホストへ書込み完了
① ホストから Write
(N) Write をｻﾌﾞｼｽﾃﾑ２へ転送
(N+1) 書込み完了
記憶ｻﾌﾞｼｽﾃﾑ１
記憶ｻﾌﾞｼｽﾃﾑ２
ﾘｰﾄ機能により
ホストからのﾃﾞｰﾀ受領順で
記憶ｻﾌﾞｼｽﾃﾑ２へ格納
ﾃﾞｰﾀ転送順序
Data #1
Data #2
Data #4
Data #3

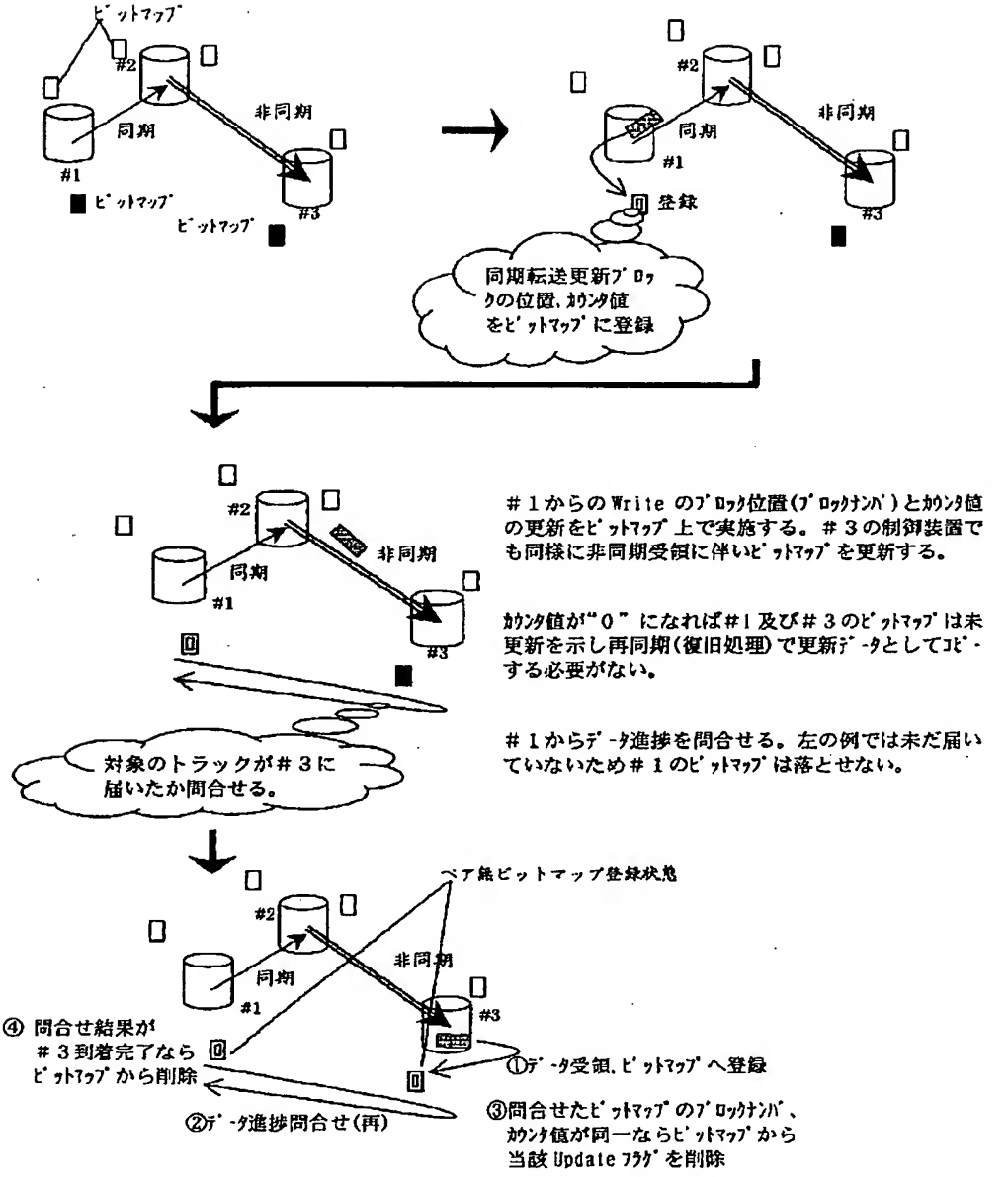
【図７】


障害時の運用と効果
#2
同期
非同期
#1
#3
ﾃﾞｰﾀ進捗問合せ
正常時
障害時
ﾘｶﾊﾞﾘ

【図８】


障害時の運用と効果
#2
ﾃﾞｰﾀ進捗問合せ
同期
非同期
#1
#3
正常時
障害時
ﾘｶﾊﾞﾘ

【図１１】



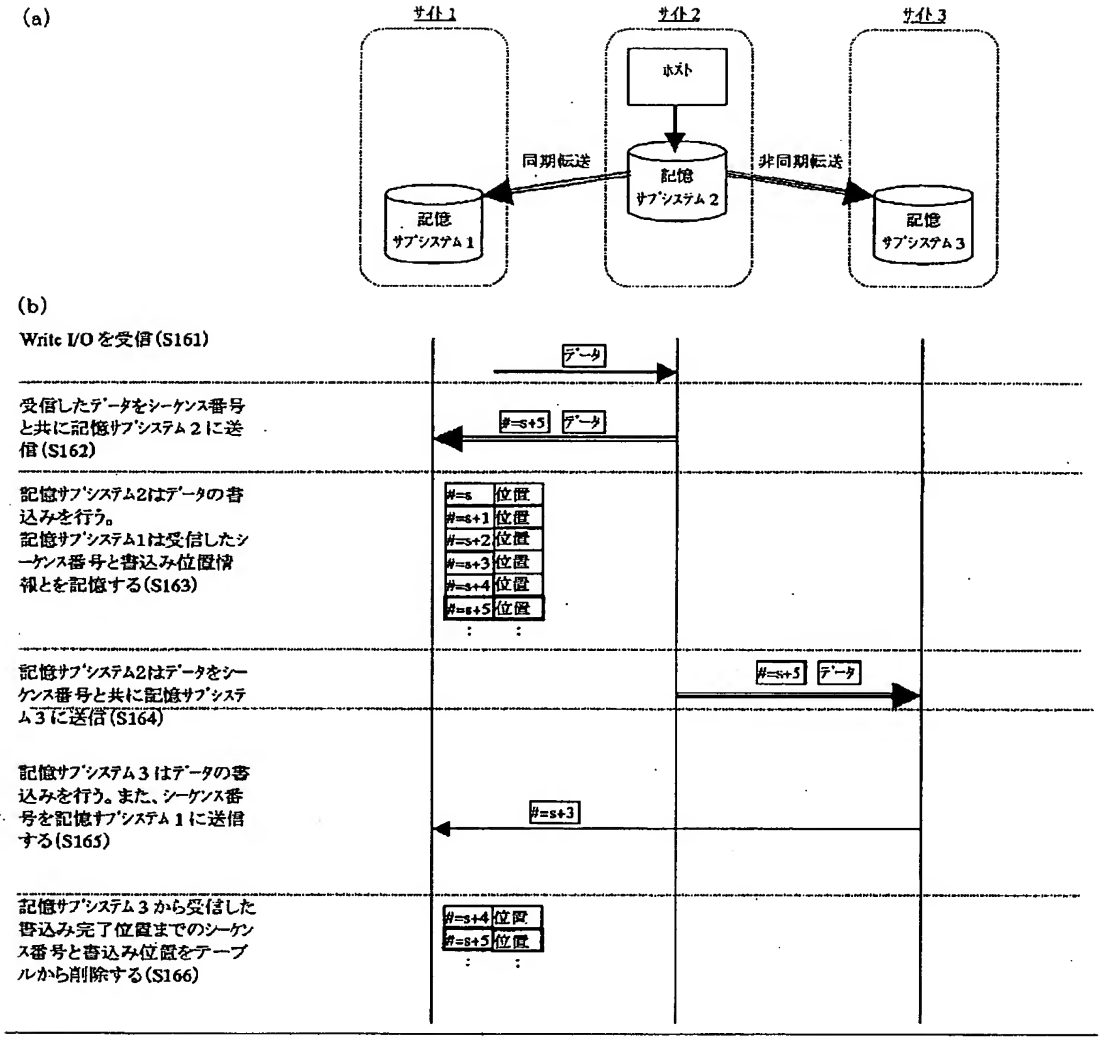
【図１２】

ブロック番号56	ブロック番号57	ブロック番号58	ブロック番号59

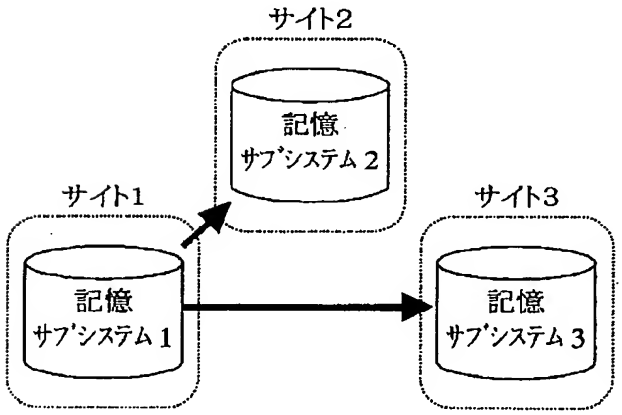
【図１４】

ブロック番号 シーケンス番号	データ	ブロック番号 シーケンス番号	データ	ブロック番号 シーケンス番号	データ

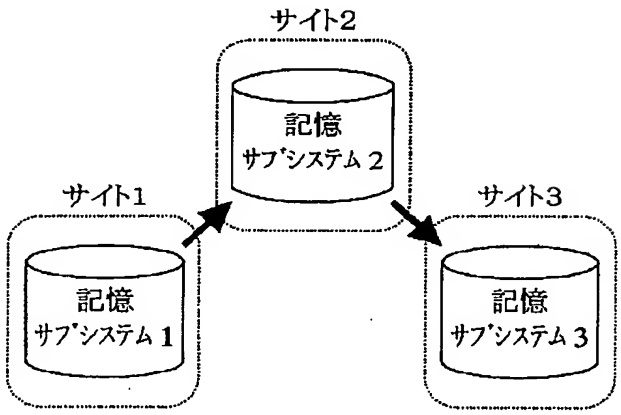
【図１６】


(a)
サイト1
サイト2
サイト3
ホスト
同期転送
非同期転送
記憶
サブシステム1
記憶
サブシステム2
記憶
サブシステム3
(b)
Write I/Oを受信(S161)
データ
受信したデータをシーケンス番号と共に記憶サブシステム2に送信(S162)
#=s+5
データ
記憶サブシステム2はデータの書込みを行う。記憶サブシステム1は受信したシーケンス番号と書込み位置情報とを記憶する(S163)
#=s
#=s+1
#=s+2
#=s+3
#=s+4
#=s+5
位置
記憶サブシステム2はデータをシーケンス番号と共に記憶サブシステム3に送信(S164)
記憶サブシステム3はデータの書込みを行う。また、シーケンス番号を記憶サブシステム1に送信する(S165)
記憶サブシステム3から受信した書込み完了位置までのシーケンス番号と書込み位置をテーブルから削除する(S166)

【図１８】

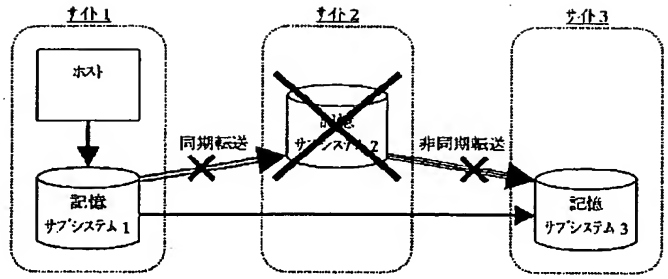

サイト2
記憶
サブシステム2
サイト1
サイト3
記憶
サブシステム1
記憶
サブシステム3

【図１９】

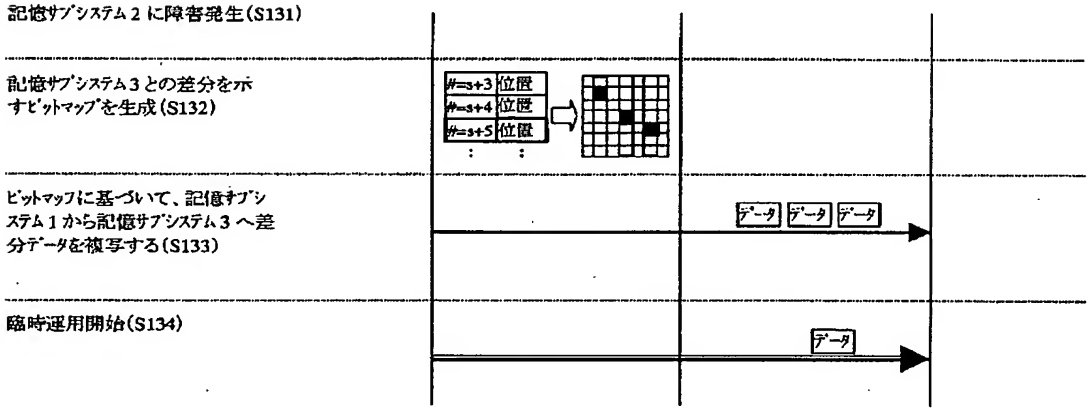

サイト2
記憶
サブシステム2
サイト1
サイト3
記憶
サブシステム1
記憶
サブシステム3

【図１７】

(a)

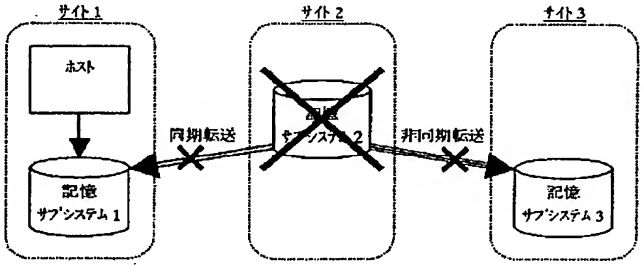

サイト１
サイト２
サイト３
ホスト
同期転送
非同期転送
記憶
サブシステム１
記憶
サブシステム３

(b)

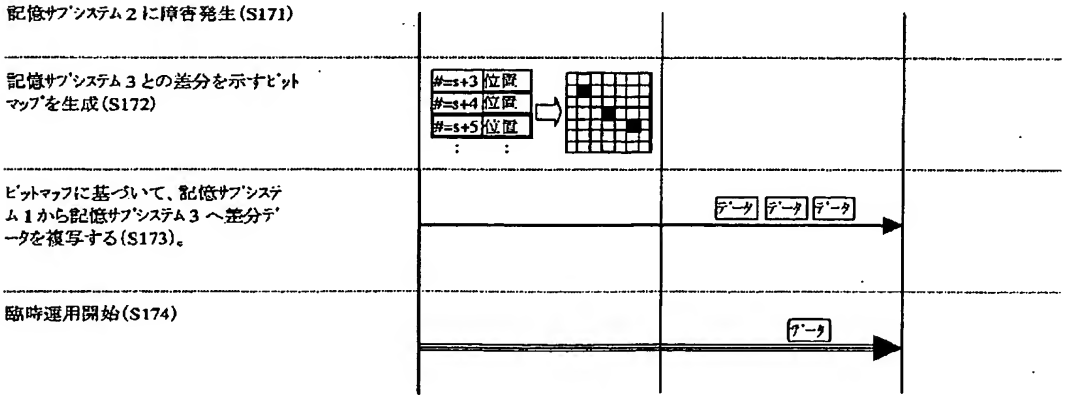

記憶サブシステム２に障害発生(S131)
記憶サブシステム３との差分を示すビットマップを生成(S132)
#=s+3 位置
#=s+4 位置
#=s+5 位置
ビットマップに基づいて、記憶サブシステム１から記憶サブシステム３へ差分データを複写する(S133)
データ
データ
データ
臨時運用開始(S134)
データ

【図２１】

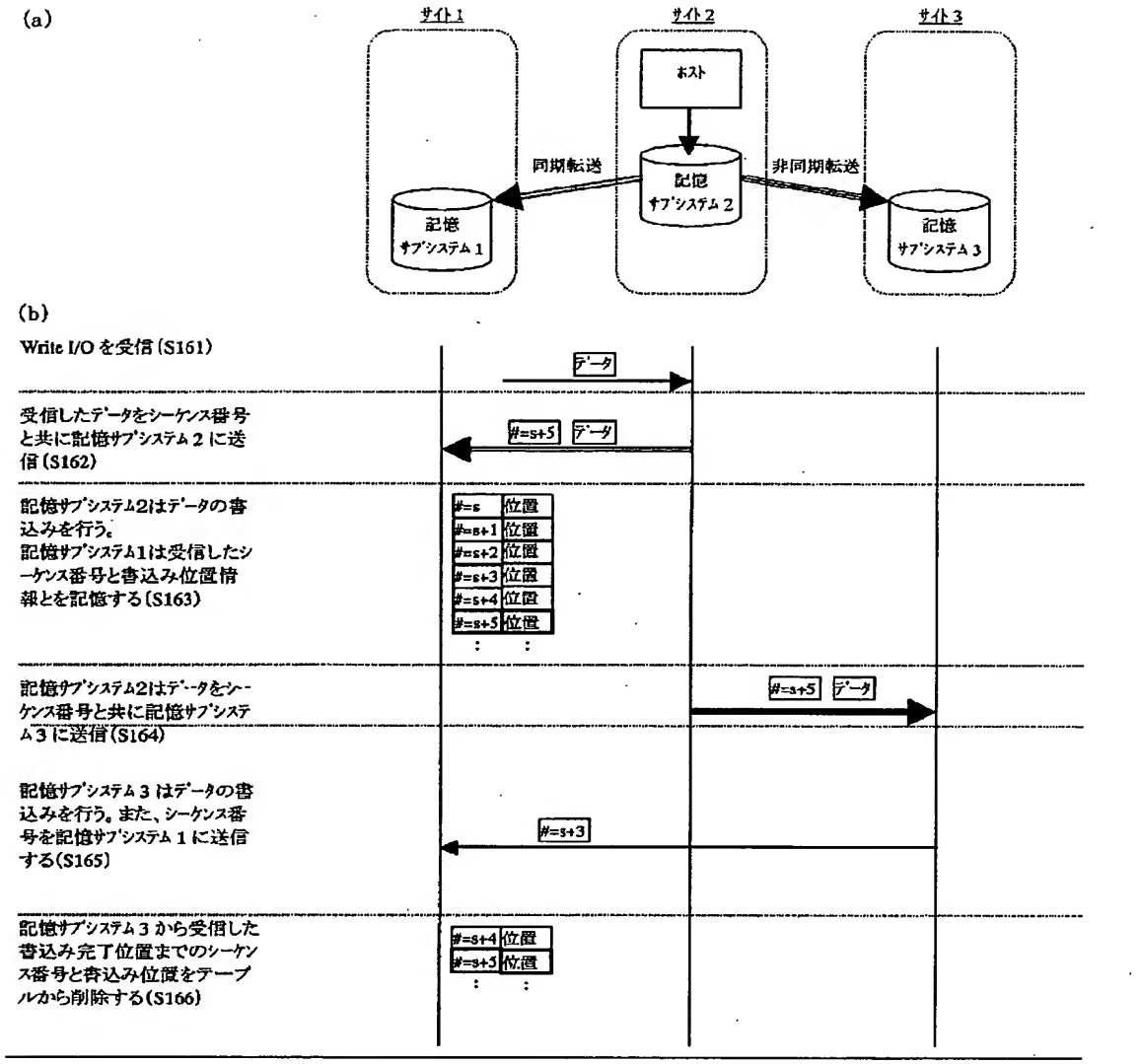
(a)


サイト１
サイト２
サイト３
ホスト
同期転送
非同期転送
記憶
サブシステム１
記憶
サブシステム３

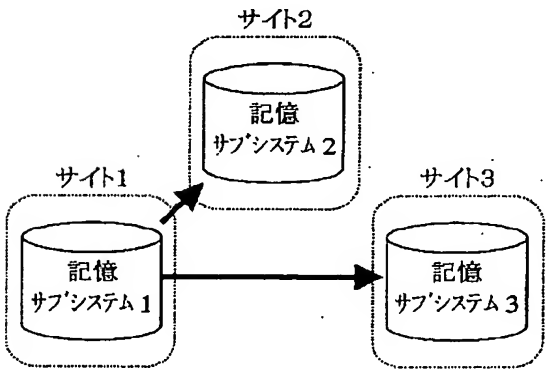
(b)


記憶サブシステム２に障害発生(S171)
記憶サブシステム３との差分を示すビットマップを生成(S172)
#=s+3 位置
#=s+4 位置
#=s+5 位置
ビットマップに基づいて、記憶サブシステム１から記憶サブシステム３へ差分データを複写する(S173)。
データ
データ
データ
臨時運用開始(S174)
データ

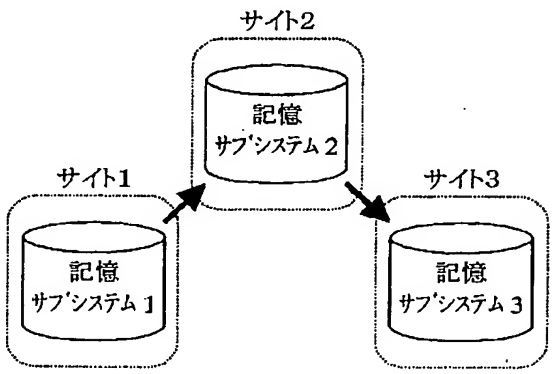
【図２０】


(a)
サイト1
サイト2
サイト3
ホスト
同期転送
非同期転送
記憶
サブシステム1
記憶
サブシステム2
記憶
サブシステム3
(b)
Write I/O を受信(S161)
データ
受信したデータをシーケンス番号と共に記憶サブシステム2に送信(S162)
#=s+5
データ
記憶サブシステム2はデータの書込みを行う。
記憶サブシステム1は受信したシーケンス番号と書込み位置情報とを記憶する(S163)
#=s
#=s+1
#=s+2
#=s+3
#=s+4
#=s+5
位置
記憶サブシステム2はデータをシーケンス番号と共に記憶サブシステム3に送信(S164)
#=s+5
データ
記憶サブシステム3はデータの書込みを行う。また、シーケンス番号を記憶サブシステム1に送信する(S165)
#=s+3
記憶サブシステム3から受信した書込み完了位置までのシーケンス番号と書込み位置をテーブルから削除する(S166)
#=s+4
#=s+5
位置

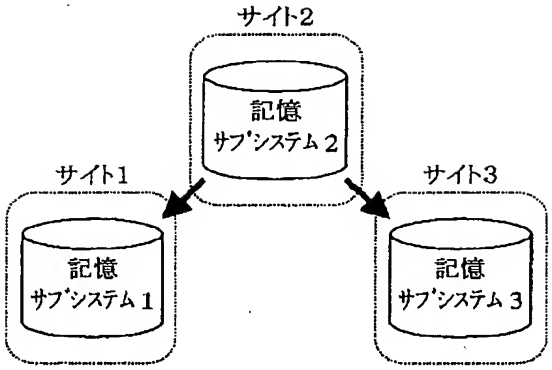
【図２２】


サイト2
記憶
サブシステム2
サイト1
サイト3
記憶
サブシステム1
記憶
サブシステム3

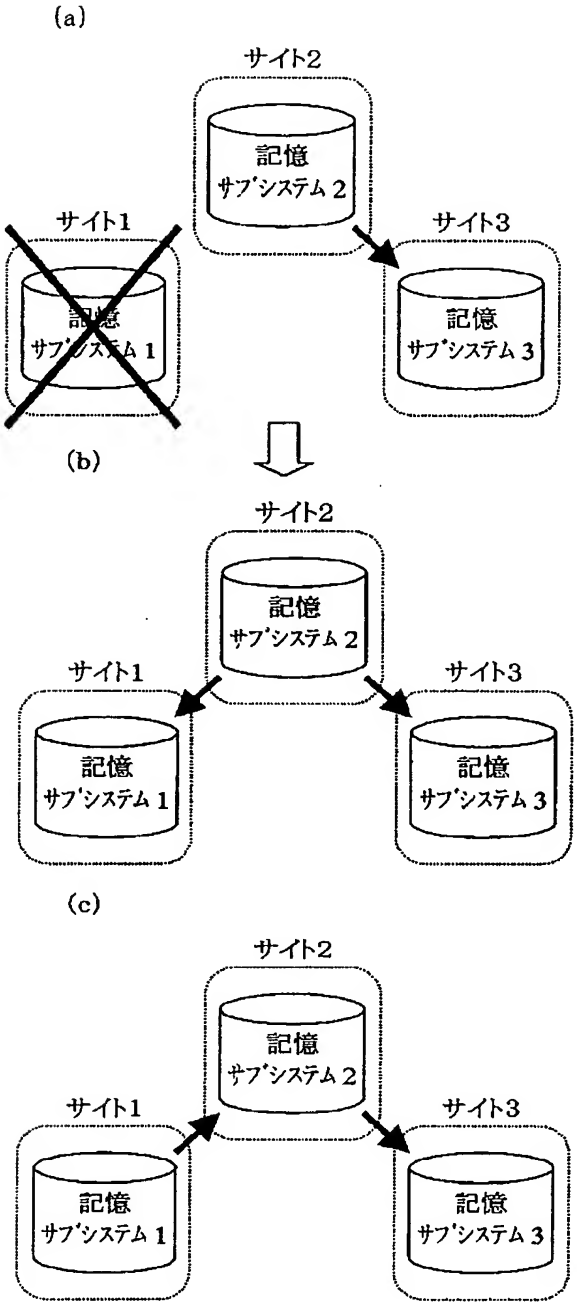
【図２３】


サイト2
記憶
サブシステム2
サイト1
サイト3
記憶
サブシステム1
記憶
サブシステム3

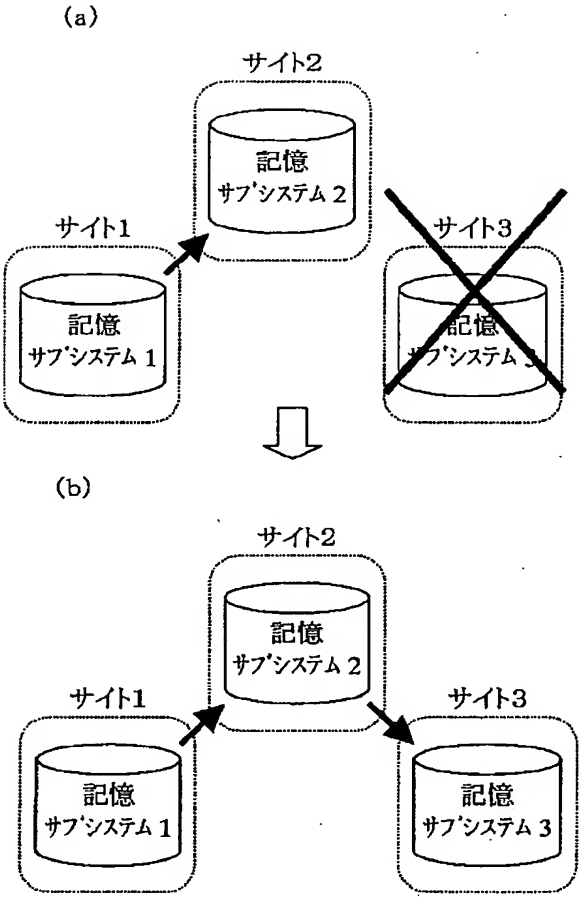
【図２４】


サイト2
記憶
サブシステム２
サイト1
記憶
サブシステム１
サイト3
記憶
サブシステム３

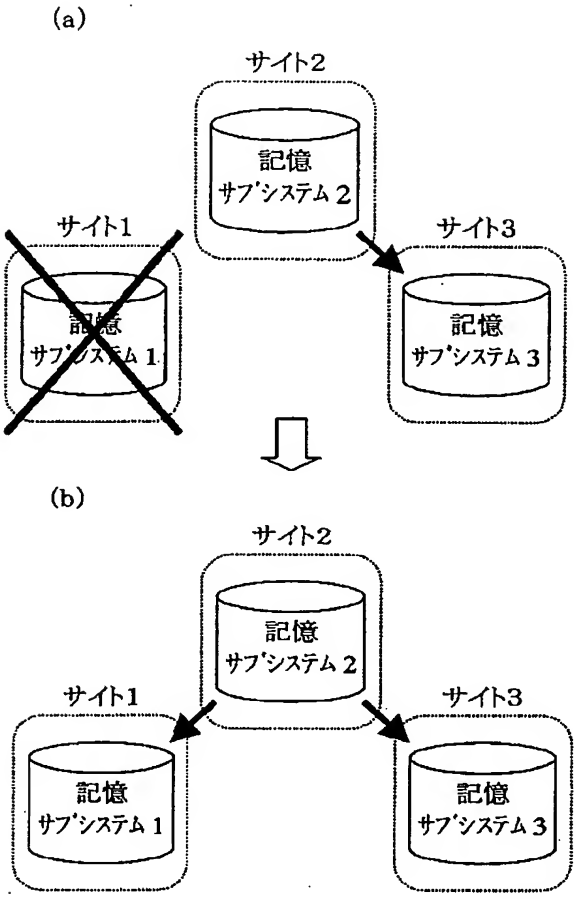
【図２５】


(a)
サイト2
記憶
サブシステム２
サイト1
記憶
サブシステム１
サイト3
記憶
サブシステム３
(b)
サイト2
記憶
サブシステム２
サイト1
記憶
サブシステム１
サイト3
記憶
サブシステム３
(c)
サイト2
記憶
サブシステム２
サイト1
記憶
サブシステム１
サイト3
記憶
サブシステム３

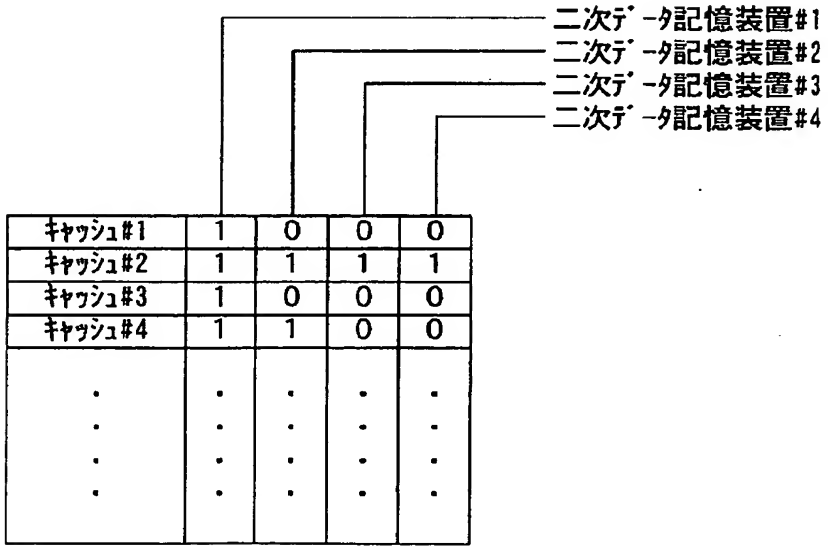
【図２６】


(a)
サイト2
記憶
サブシステム 2
サイト1
記憶
サブシステム 1
サイト3
記憶
サブシステム 3
(b)
サイト2
記憶
サブシステム 2
サイト1
記憶
サブシステム 1
サイト3
記憶
サブシステム 3

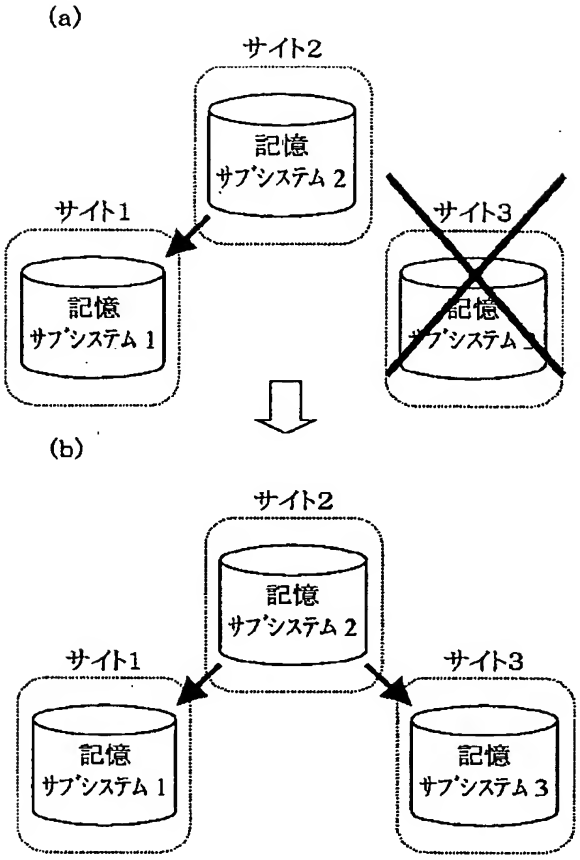
【図２７】


(a)
サイト2
記憶
サブシステム 2
サイト1
サイト3
記憶
サブシステム 3
(b)
サイト2
記憶
サブシステム 2
サイト1
記憶
サブシステム 1
サイト3
記憶
サブシステム 3

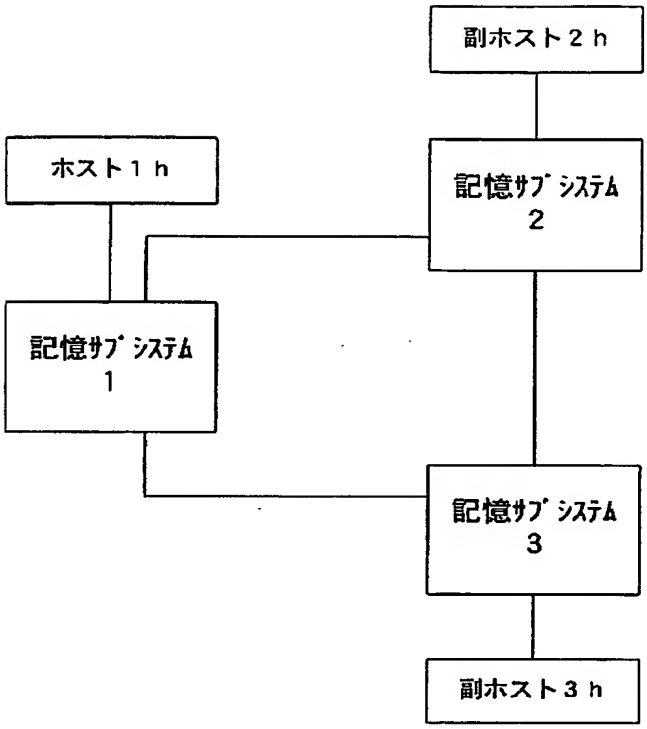
【図３０】


二次ﾃﾞｰﾀ記憶装置#1
二次ﾃﾞｰﾀ記憶装置#2
二次ﾃﾞｰﾀ記憶装置#3
二次ﾃﾞｰﾀ記憶装置#4
ｷｬｯｼｭ#1 1 0 0 0
ｷｬｯｼｭ#2 1 1 1 1
ｷｬｯｼｭ#3 1 0 0 0
ｷｬｯｼｭ#4 1 1 0 0

【図２８】



【図２９】



フロントページの続き

(72)発明者　尾形　幹人
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　岡見　吉規
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　檜垣　誠一
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　安部井　大
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　木城　茂
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

Ｆターム(参考)　5B065 BA01 CE01 EA35
5B082 DE03 GA04